

# Arithmetic Swap Notebook

Set up the initial libraries and dependant functions.

```
#Installation example:
#install.packages('hexbin', repos='http://cran.us.r-project.org')
#Knit doesn't seem to be working in RStudio, R command:
#require("knitr")
#opts_knit$set(root.dir = "/data/anon/SemanticTransformation")
#rmarkdown::render("./RSource/TransformationComparison.Rmd", params = list())

#Note: Either before this or in this R script, remove cases where the
#transformation = original (Only really need to check in the swap)
setwd("/data/anon/SemanticTransformation/")

library(lmerTest)
library(car)
library(sqldf)
library(effsize)
library(GGally)
library(compiler)
library(dplyr)
library(MuMIn) #Mixed effects R squared
library(Hmisc)
library(stargazer)
library(xtable)

source("./RSource/RegressionHelper.R")
source("./RSource/ComparisonTestHelper.R")
source("./RSource/FileLoadHelper.R")
source("./RSource/GitInfoTableFunctions.R")
source("./RSource/TransSummaryFunctions.R")
source("./RSource/TSumCompareFunctions.R")
source("./RSource/ColorBlind.R")
source("./RSource/ResultPrinter.R")
source("./RSource/LMDiffModels.R")
```

‘+’/’\*’ Swaping Results

```
setwd("/data/anon/SemanticTransformation/")

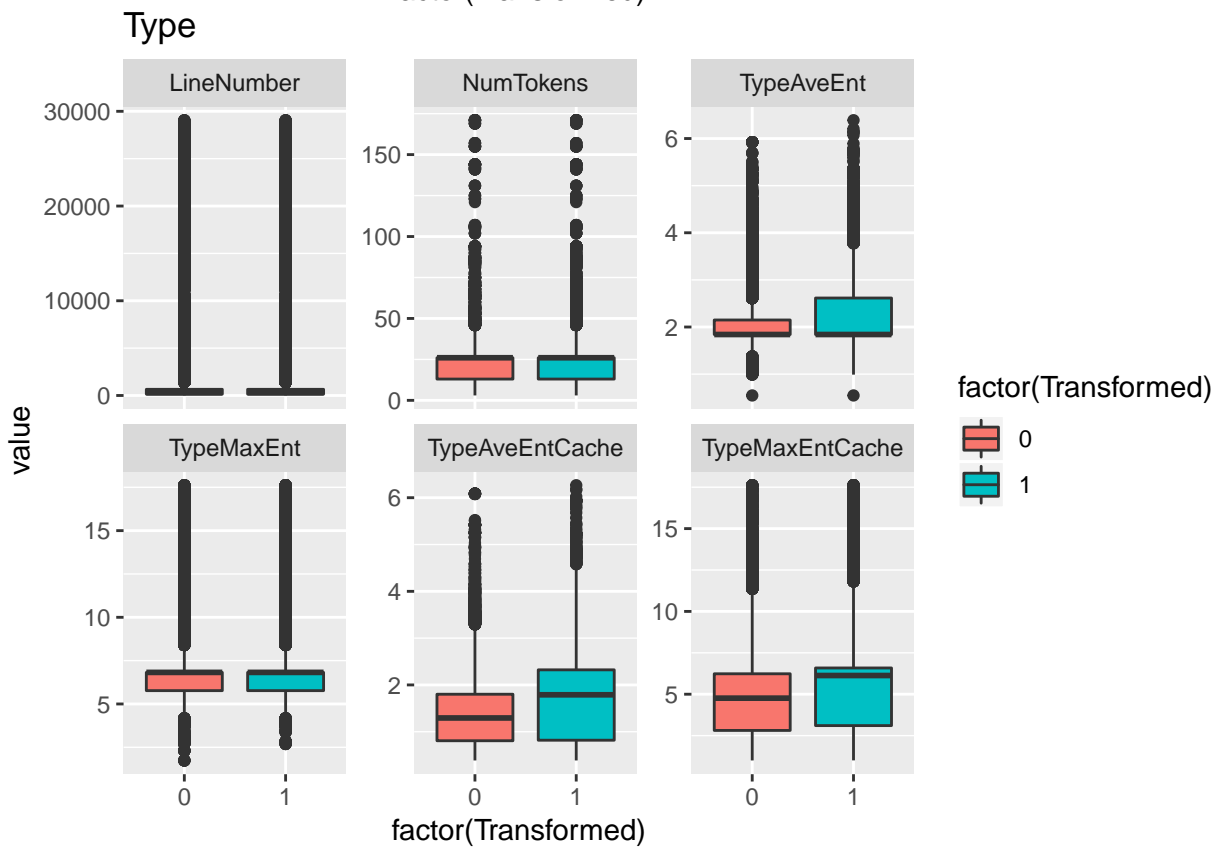
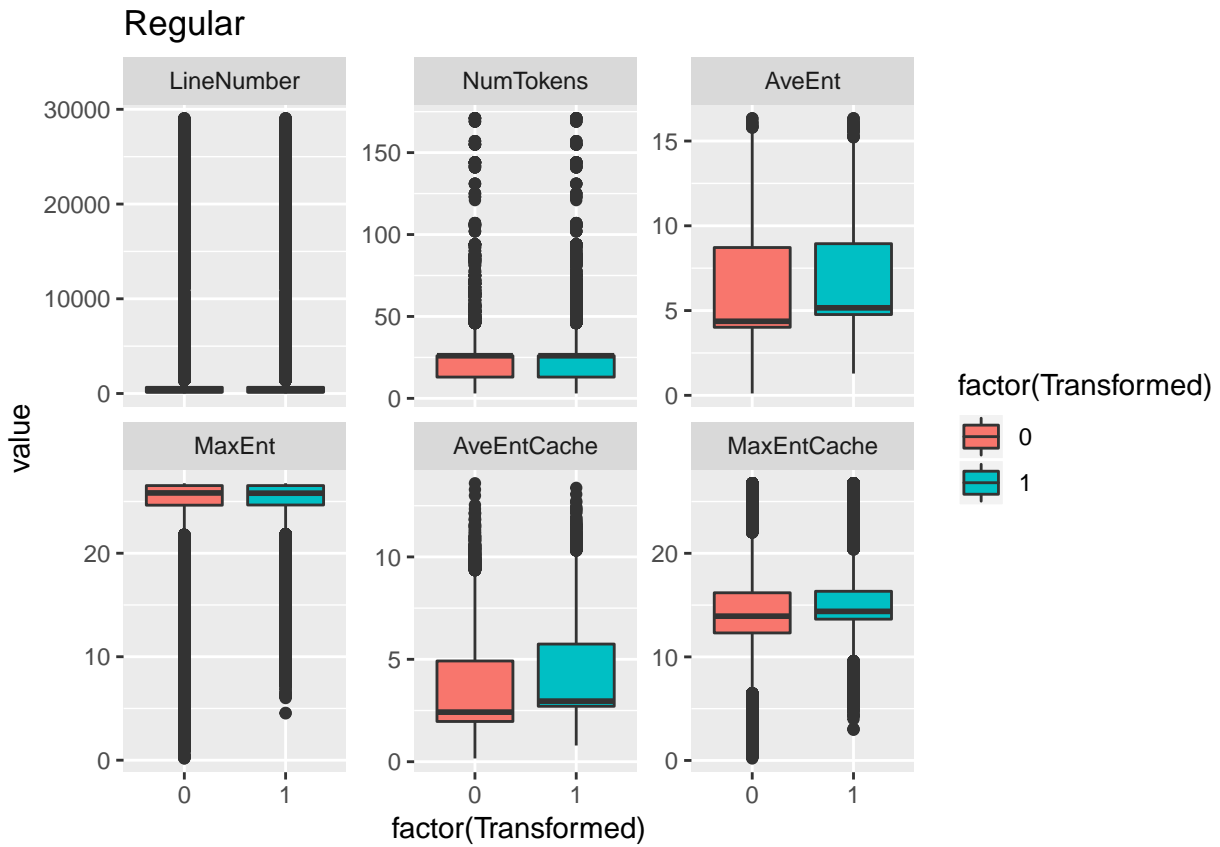
#Plot results on everything (for paired table)
#dsp <- compareDepthSummary("swap.csv", "Swap", "SAME", FALSE)
#Plot for the Large dataset
dsp <- compareDepthSummary("swap_topstarred.csv", "ArithSwapTop", "SAME", FALSE)

## [[1]]
## [1] "5"    "Bool"

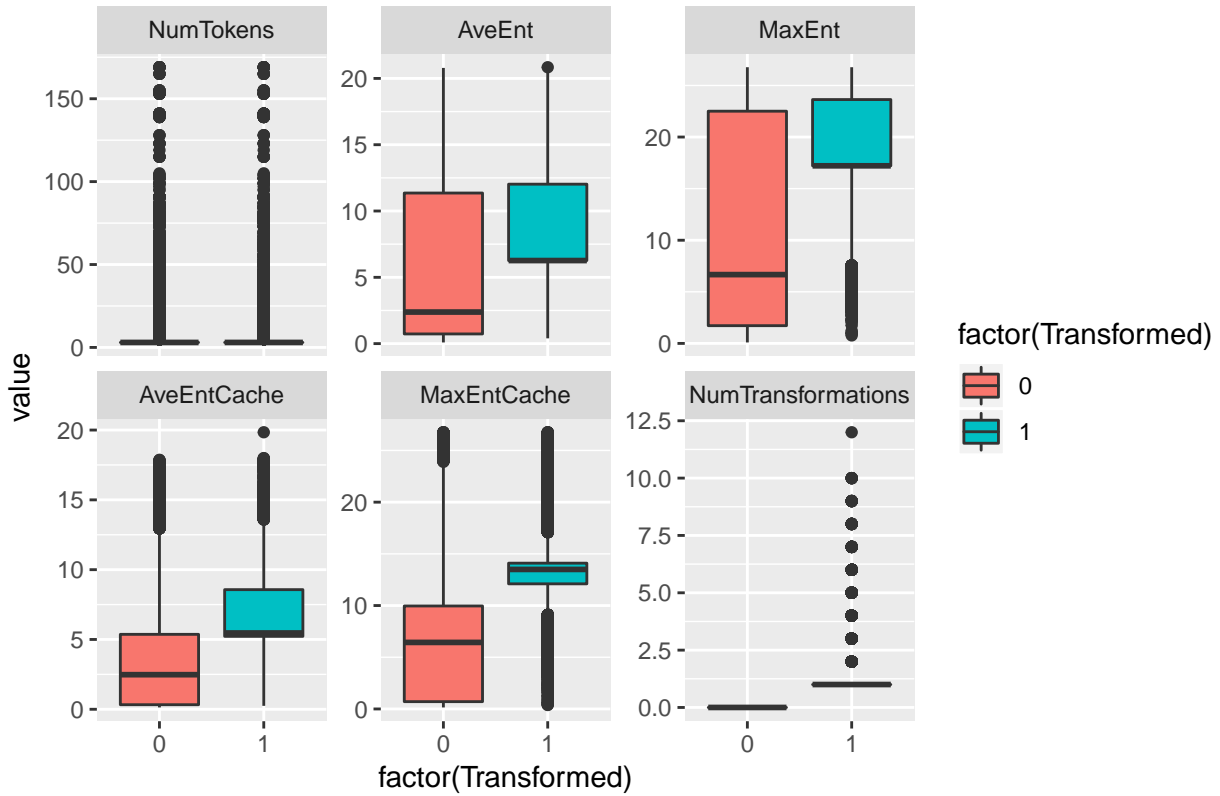
## Loading required package: tcltk

## [1] "TransId"
## [2] "Filepath"
## [3] "LineNumber"
## [4] "NumTokens"
```

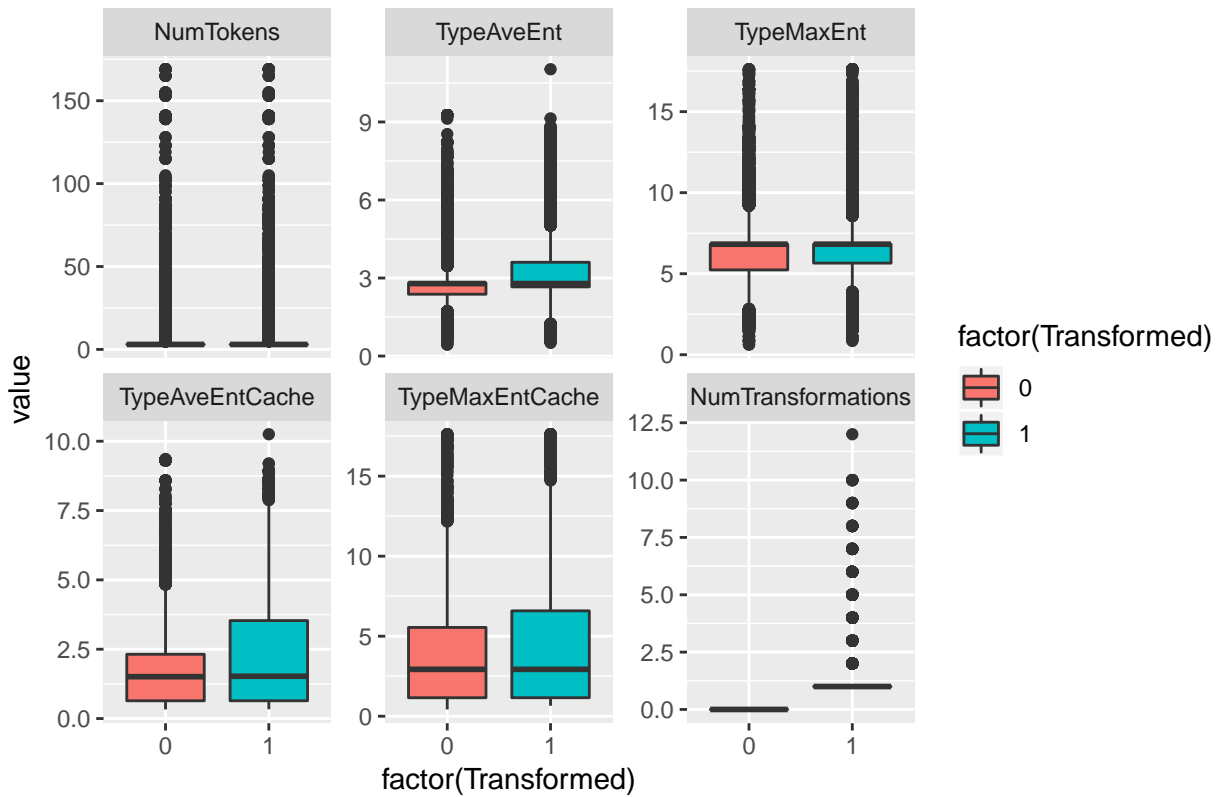
```
## [5] "Transformed"
## [6] "Source"
## [7] "CleanLexerNumTokens"
## [8] "CleanLexerSource"
## [9] "AveEnt"
## [10] "MaxEnt"
## [11] "AveEntCache"
## [12] "MaxEntCache"
## [13] "AveEntRev"
## [14] "MaxEntRev"
## [15] "TypeSource"
## [16] "TypeNumTokens"
## [17] "TypeAveEnt"
## [18] "TypeMaxEnt"
## [19] "TypeAveEntCache"
## [20] "TypeMaxEntCache"
## [21] "Depth"
## [22] "Expression"
## [23] "ExpressionNumTokens"
## [24] "ExpressionCleanLexerSource"
## [25] "ExpressionCleanLexerNumTokens"
## [26] "ExpressionForwardAverageEntropy"
## [27] "ExpressionForwardMaxEntropy"
## [28] "ExpressionForwardAverageEntropyCache"
## [29] "ExpressionForwardMaxEntropyCache"
## [30] "TypeExpression"
## [31] "ExpressionTypeNumTokens"
## [32] "ExpressionTypeAverageEntropy"
## [33] "ExpressionTypeMaxEntropy"
## [34] "ExpressionTypeAverageEntropyCache"
## [35] "ExpressionTypeMaxEntropyCache"
## [36] "NumTransformations"
## [37] "ParentOp"
## [38] "MostFreqOp"
## [39] "LeastFreqOp"
## [40] "MostFreqParentOp"
## [41] "ParentChildFreq"
## [42] "ParentParenChildFreq"
## [43] "PoolSize"
## [44] "TransSetNo"
## [45] "TransNo"
## [46] "Type"
## [47] "NumTypes"
## [48] "MethodName"
## [49] "rowID"
```



### Expression (Regular)



### Expression (Type)

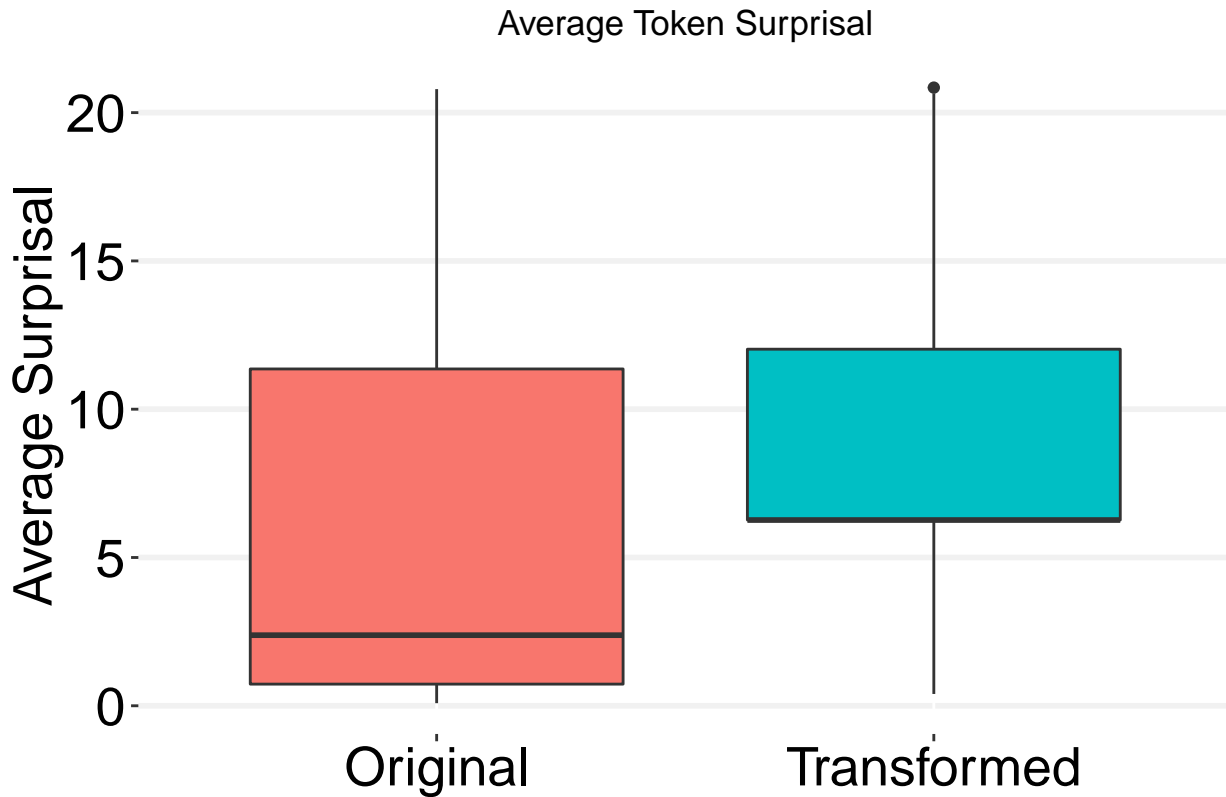


## [1] " ----- Expression Global Model ----- "

```

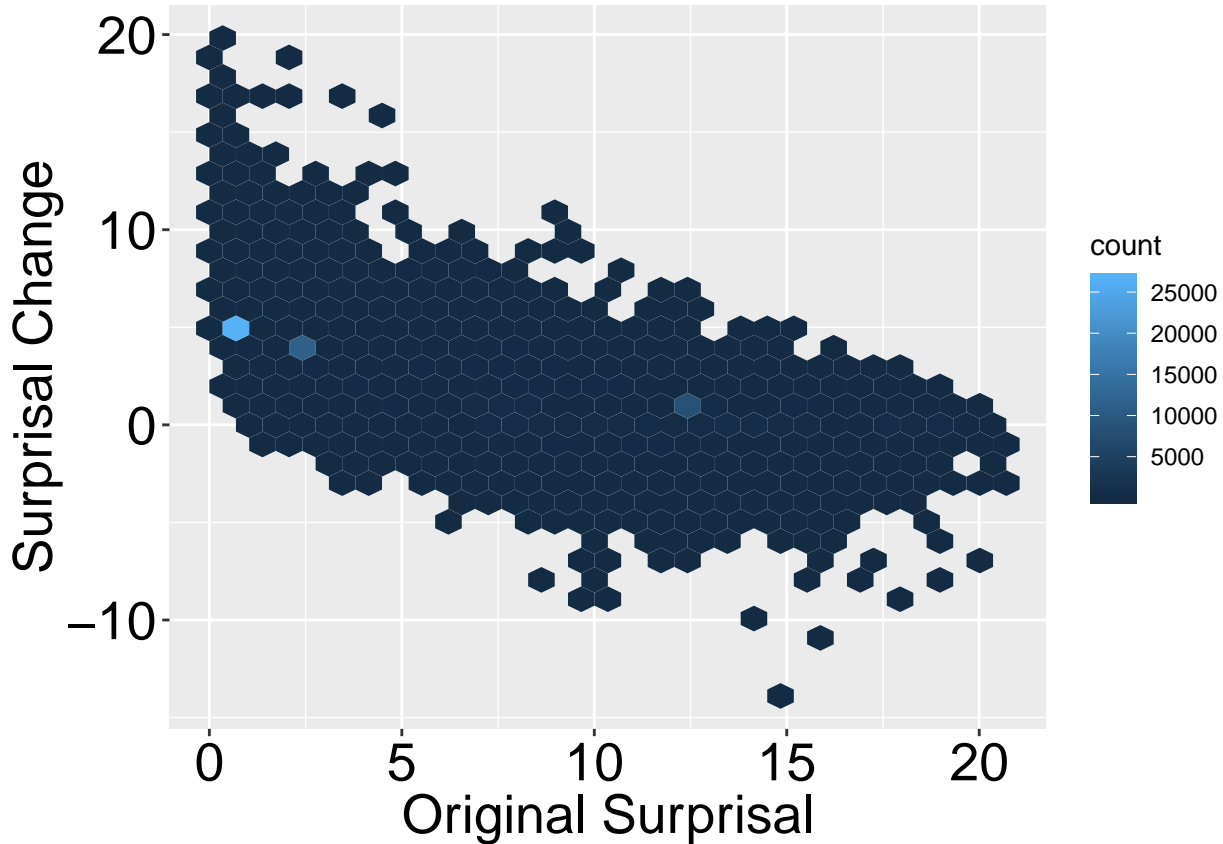
## [1] "ArithSwapTopGlobalExp Original < Transformed"
##
## Paired t-test
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## t = -326.56, df = 66765, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 99.80769 percent confidence interval:
##      -Inf -3.165802
## sample estimates:
## mean of the differences
##      -3.194075
##
## Paired t-test
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## t = -326.56, df = 66765, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99.80769 percent confidence interval:
##  -3.224416 -3.163735
## sample estimates:
## mean of the differences
##      -3.194075
##
## Warning in n1 * n2: NAs produced by integer overflow
##
## Cohen's d
##
## d estimate: -1.263802 (large)
## 95 percent confidence interval:
## inf sup
## NA NA
##
## Warning in n1 * n2: NAs produced by integer overflow
##
## Warning in n1 * n2: NAs produced by integer overflow

```



```
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 80389000, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
## 99.80769 percent confidence interval:
##      -Inf -3.12251
## sample estimates:
## (pseudo)median
##      -3.122516
##
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 80389000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
## 99.80769 percent confidence interval:
##      -3.122490 -3.122566
## sample estimates:
## (pseudo)median
##      -3.122516
##
##
## Cliff's Delta
##
## delta estimate: -0.4336698 (medium)
```

```
## 95 percent confidence interval:
##      inf      sup
## -0.4396828 -0.4276180
```

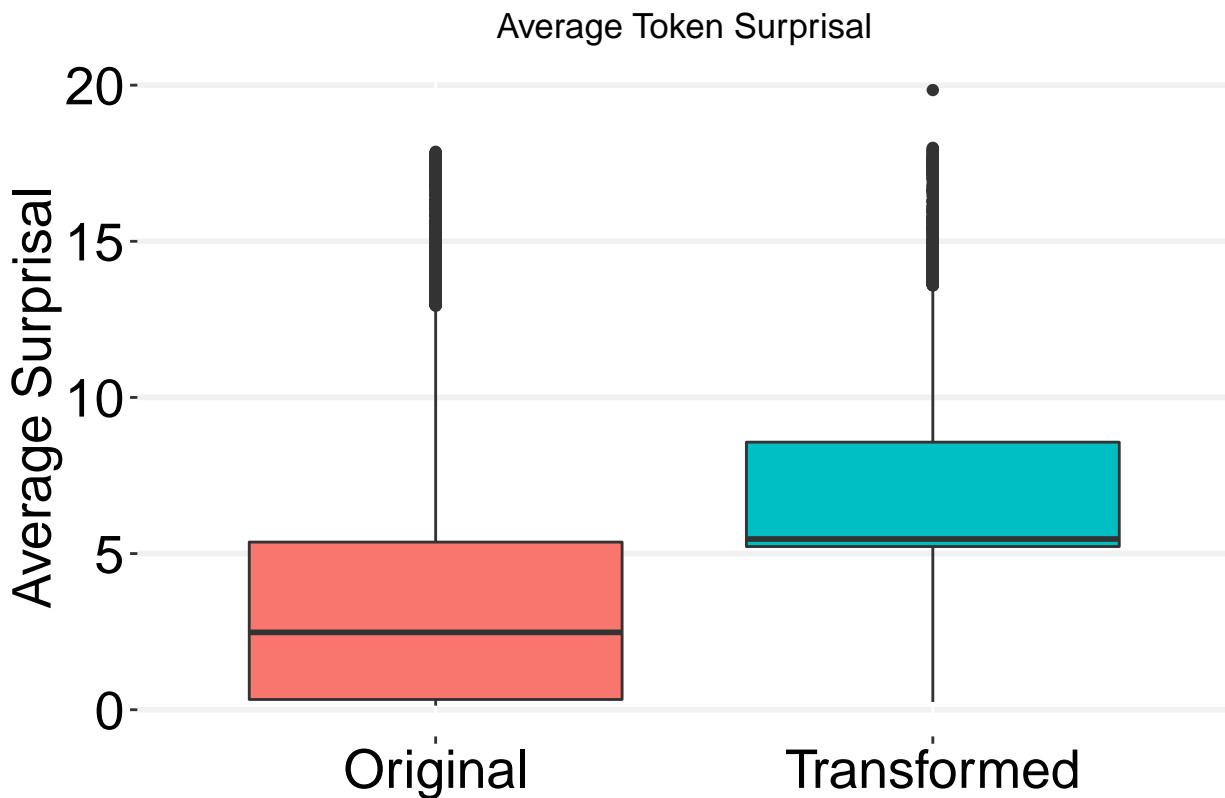


```
## [1] "----- Expression Cache Model -----"
## [1] "ArithSwapTopCacheExp Original < Transformed"
##
## Paired t-test
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## t = -359.85, df = 66765, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 99.80769 percent confidence interval:
##      -Inf -3.331353
## sample estimates:
## mean of the differences
##      -3.35833
##
## Paired t-test
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## t = -359.85, df = 66765, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99.80769 percent confidence interval:
##      -3.387279 -3.329381
## sample estimates:
```

```

## mean of the differences
##          -3.35833
## Warning in n1 * n2: NAs produced by integer overflow
##
## Cohen's d
##
## d estimate: -1.392673 (large)
## 95 percent confidence interval:
## inf sup
## NA NA
## Warning in n1 * n2: NAs produced by integer overflow
## Warning in n1 * n2: NAs produced by integer overflow

```



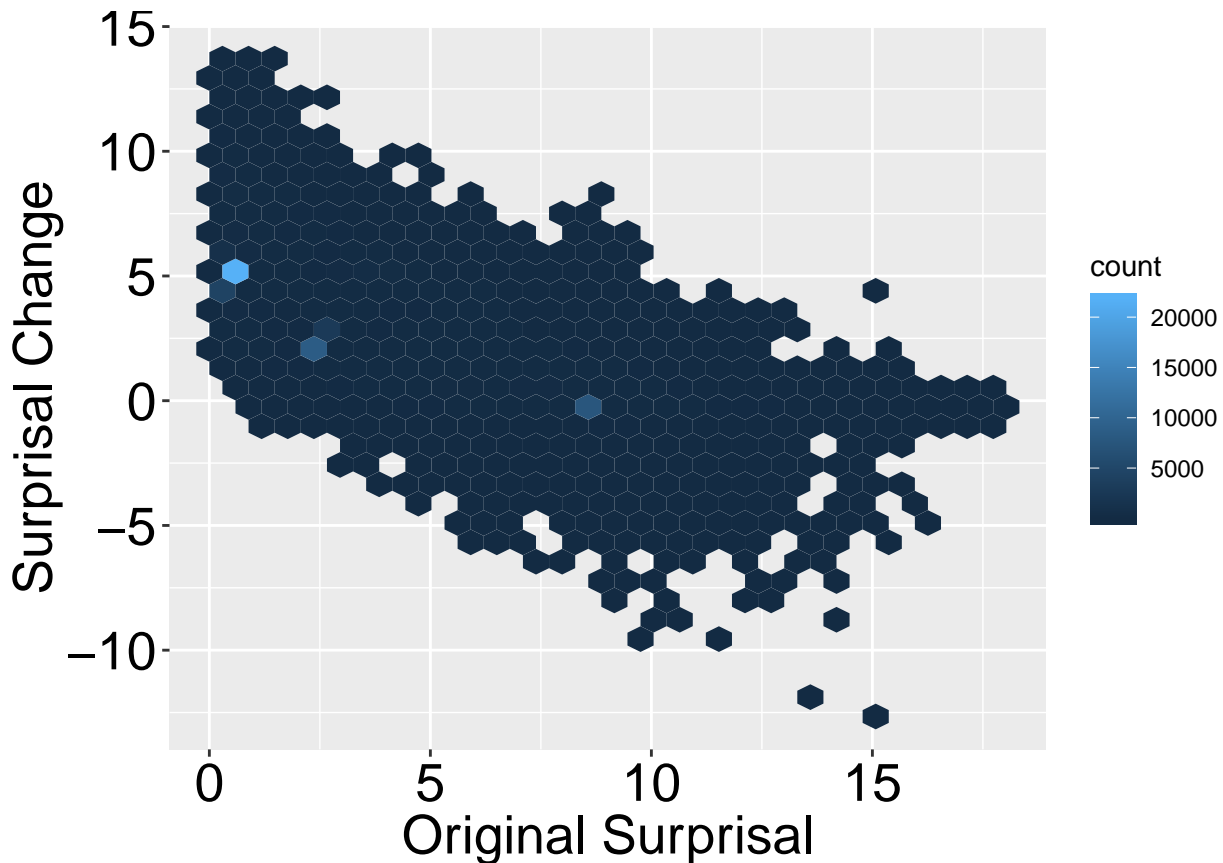
```

##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 49168000, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
## 99.80769 percent confidence interval:
##      -Inf -3.49498
## sample estimates:
## (pseudo)median
##      -3.513311
##
##

```



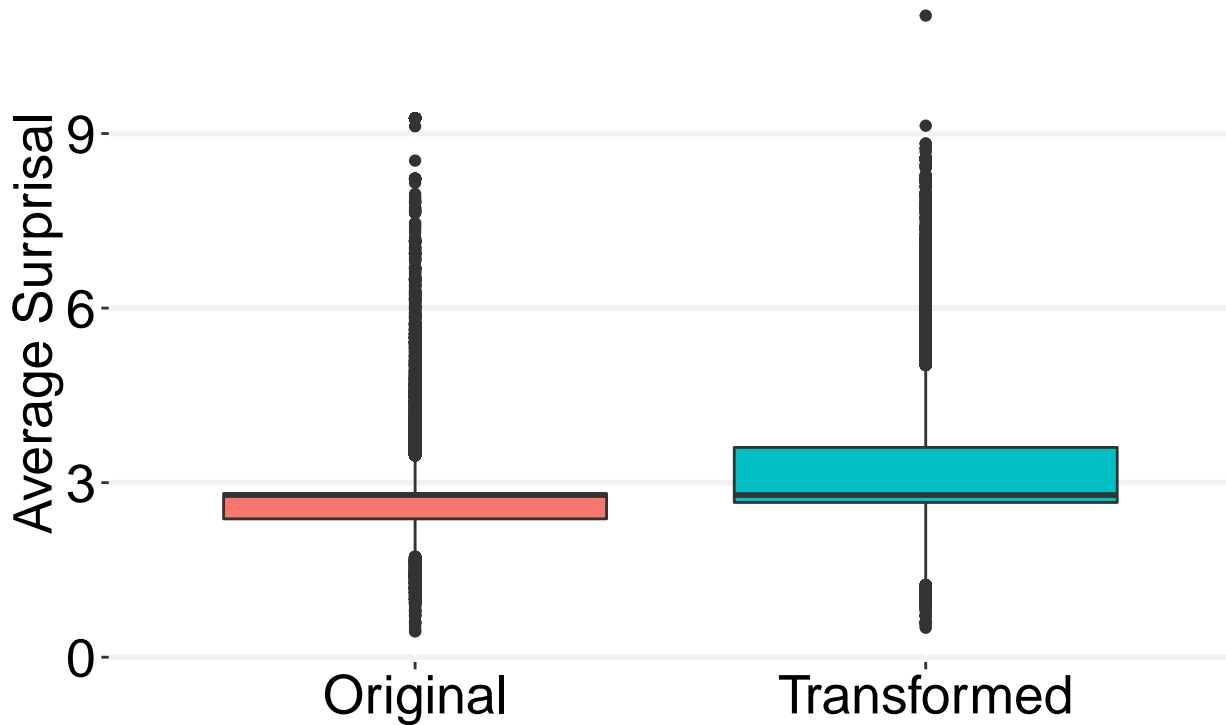
```
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 49168000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
## 99.80769 percent confidence interval:
## -3.533199 -3.492738
## sample estimates:
## (pseudo)median
## -3.513311
##
##
## Cliff's Delta
##
## delta estimate: -0.6033144 (large)
## 95 percent confidence interval:
## inf sup
## -0.6085663 -0.5980097
```



```
## [1] " ----- Expression Global Type Model ----- "
## [1] "ArithSwapTopGlobalTypeExp Original < Transformed"
##
## Paired t-test
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## t = -108.4, df = 66765, p-value < 2.2e-16
```

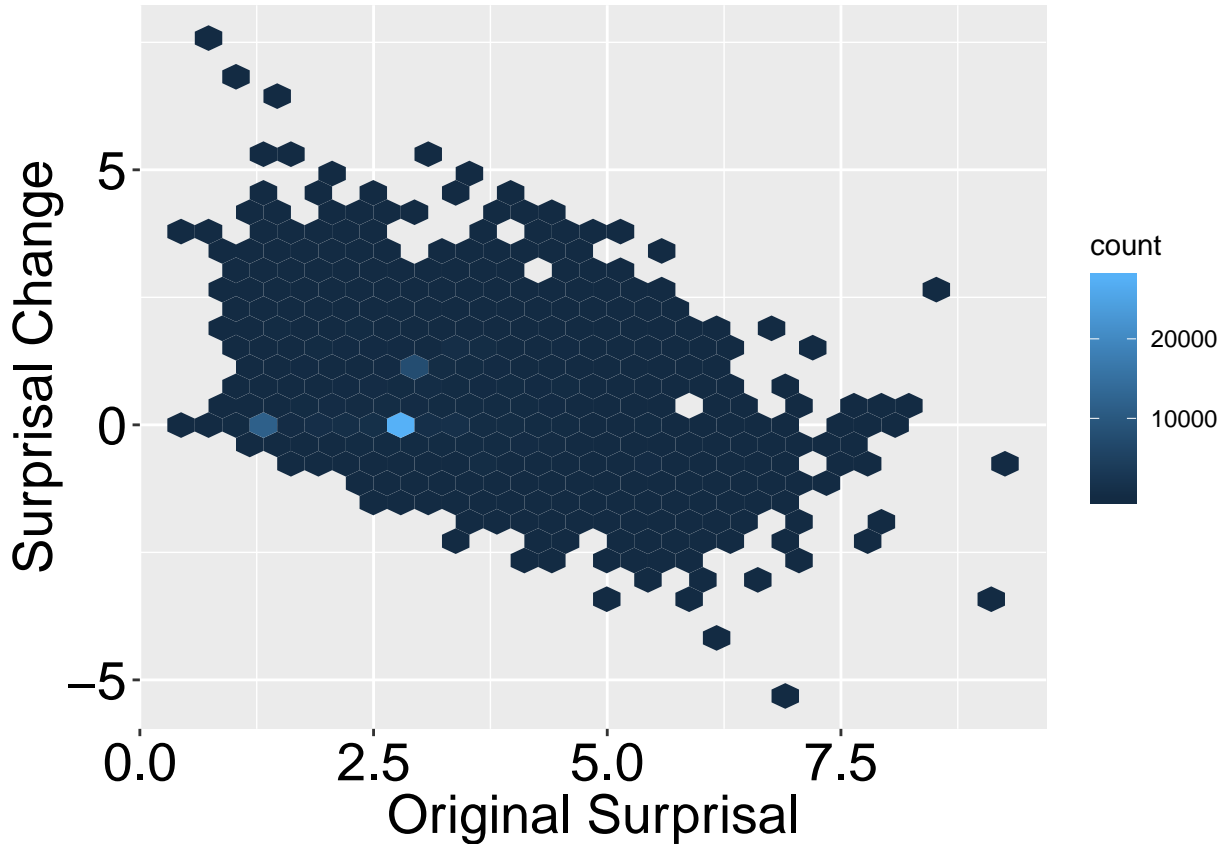
```
## alternative hypothesis: true difference in means is less than 0
## 99.80769 percent confidence interval:
##      -Inf -0.2486528
## sample estimates:
## mean of the differences
##      -0.2554652
##
##
## Paired t-test
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## t = -108.4, df = 66765, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99.80769 percent confidence interval:
## -0.2627757 -0.2481547
## sample estimates:
## mean of the differences
##      -0.2554652
##
## Warning in n1 * n2: NAs produced by integer overflow
##
## Cohen's d
##
## d estimate: -0.419515 (small)
## 95 percent confidence interval:
## inf sup
## NA NA
##
## Warning in n1 * n2: NAs produced by integer overflow
##
## Warning in n1 * n2: NAs produced by integer overflow
```

## Average Token Surprisal



```
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 28031000, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
## 99.80769 percent confidence interval:
##      -Inf -0.7309078
## sample estimates:
## (pseudo)median
##      -0.752134
##
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 28031000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
## 99.80769 percent confidence interval:
##      -0.7711736 -0.7298772
## sample estimates:
## (pseudo)median
##      -0.752134
##
##
## Cliff's Delta
##
## delta estimate: -0.09791567 (negligible)
```

```
## 95 percent confidence interval:
##      inf      sup
## -0.10388714 -0.09193713
```



```
## [1] "----- Expression Cache Type Model ----- "
```

```
## [1] "ArithSwapTopCacheTypeExp Original < Transformed"
```

```
##
```

```
## Paired t-test
```

```
##
```

```
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
```

```
## t = -130.53, df = 66765, p-value < 2.2e-16
```

```
## alternative hypothesis: true difference in means is less than 0
```

```
## 99.80769 percent confidence interval:
```

```
##      -Inf -0.4755341
```

```
## sample estimates:
```

```
## mean of the differences
```

```
##      -0.4863034
```

```
##
```

```
##
```

```
## Paired t-test
```

```
##
```

```
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
```

```
## t = -130.53, df = 66765, p-value < 2.2e-16
```

```
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 99.80769 percent confidence interval:
```

```
## -0.4978601 -0.4747466
```

```
## sample estimates:
```

```

## mean of the differences
##           -0.4863034

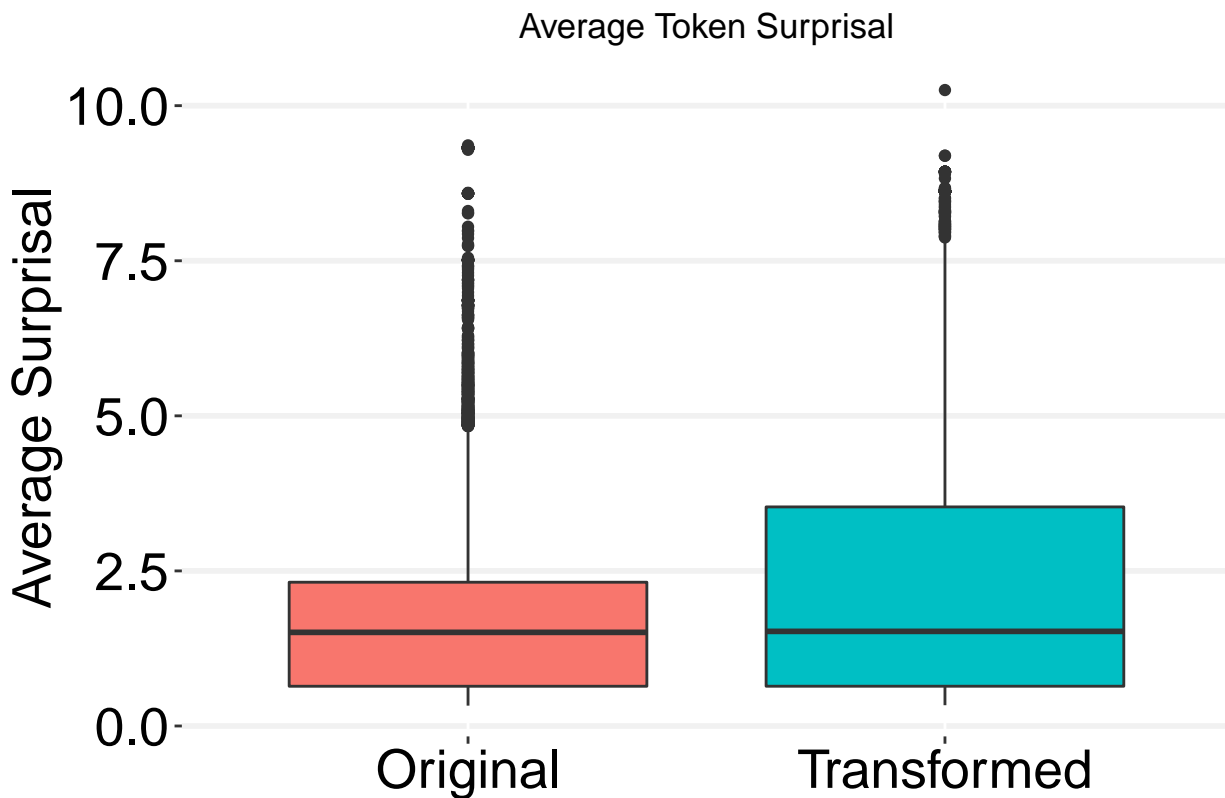
## Warning in n1 * n2: NAs produced by integer overflow

##
## Cohen's d
##
## d estimate: -0.5051653 (medium)
## 95 percent confidence interval:
## inf sup
## NA NA

## Warning in n1 * n2: NAs produced by integer overflow

## Warning in n1 * n2: NAs produced by integer overflow

```



```

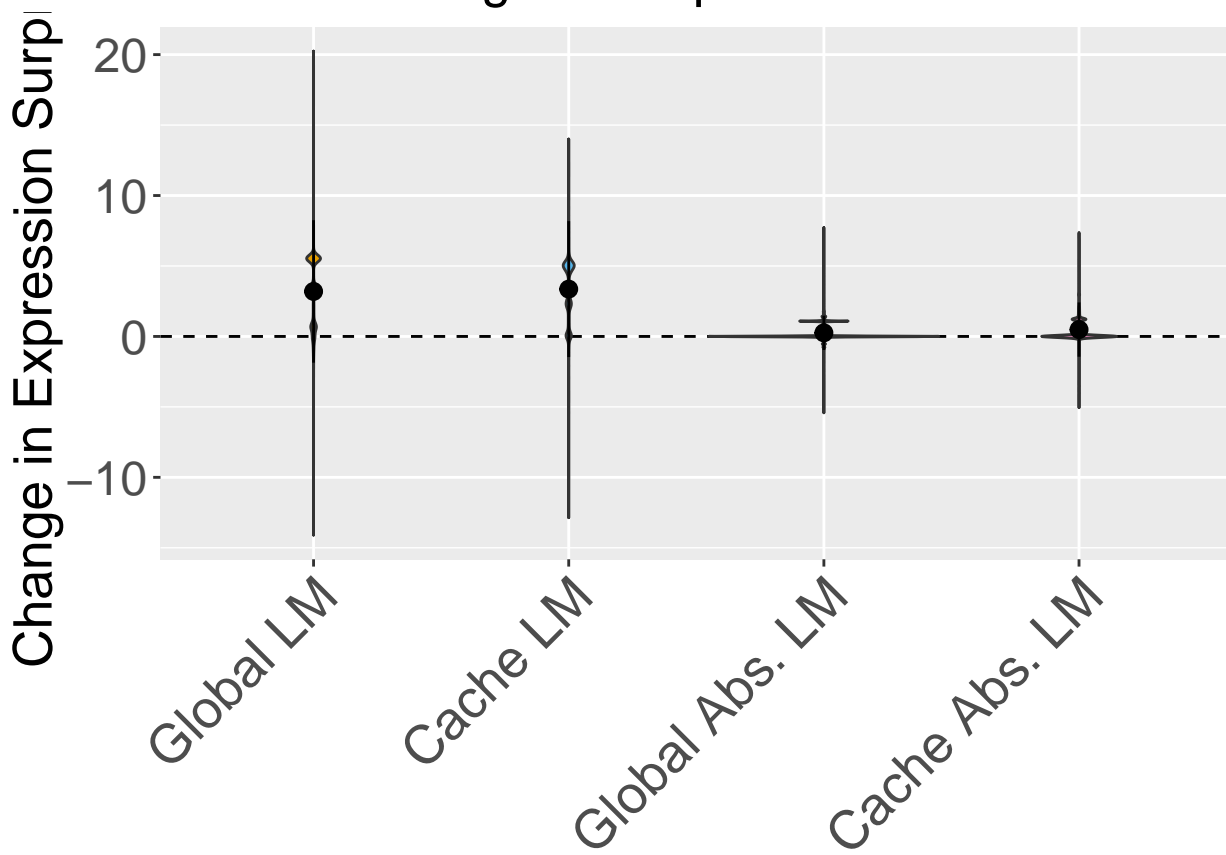
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 8350200, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
## 99.80769 percent confidence interval:
##      -Inf -1.217358
## sample estimates:
## (pseudo)median
##      -1.227661
##
##

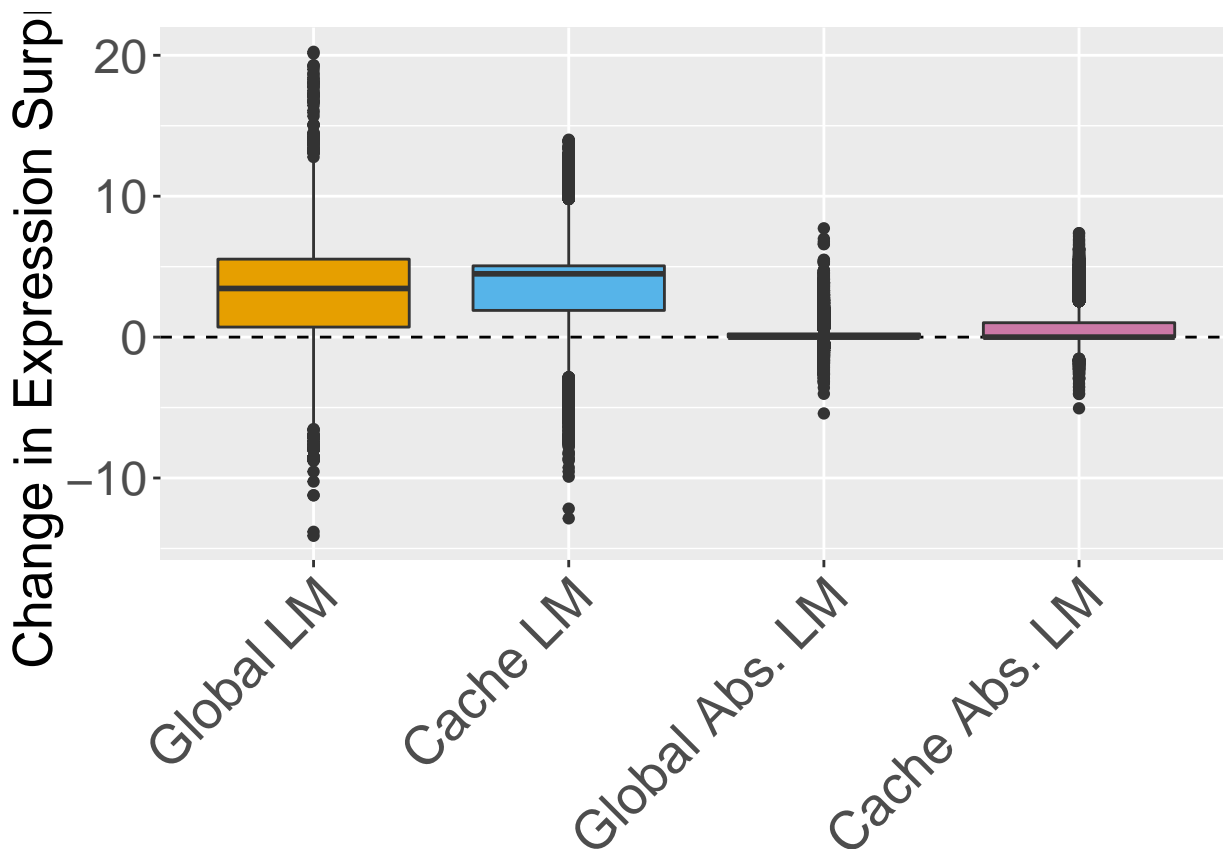
```

```

## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 8350200, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
## 99.80769 percent confidence interval:
## -1.255091 -1.217347
## sample estimates:
## (pseudo)median
## -1.227661
##
##
## Cliff's Delta
##
## delta estimate: -0.1262627 (negligible)
## 95 percent confidence interval:
## inf sup
## -0.1324445 -0.1200710
## [1] "Binary differences"
##
## FALSE TRUE
## 6333 60433
##
## FALSE TRUE
## 3292 63474
##
## FALSE TRUE
## 4760 62006
##
## FALSE TRUE
## 2111 64655
##
## No id variables; using all as measure variables
## Warning: Ignoring unknown parameters: mult

```





```

setwd("/data/anon/SemanticTransformation/")

#Plot results on everything (for paired table)
#dsp <- compareDepthSummary("swap.csv", "Swap", "SAME", FALSE)
#Plot for the Large dataset
#Give some sense of robustness to n for these values...
#Also remove multiple transform lines?
dspFiltered <- generateFilteredResults(dsp, "Arith", "ArithSwapTopFiltered100", 100)

## [1] " ----- Expression Global Model ----- "
## [1] "ArithSwapTopFiltered100GlobalExp Original < Transformed"
##
## Paired t-test
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## t = -61.445, df = 20828, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 99.80769 percent confidence interval:
##      -Inf -0.9955927
## sample estimates:
## mean of the differences
##      -1.044745
##
## Paired t-test
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt

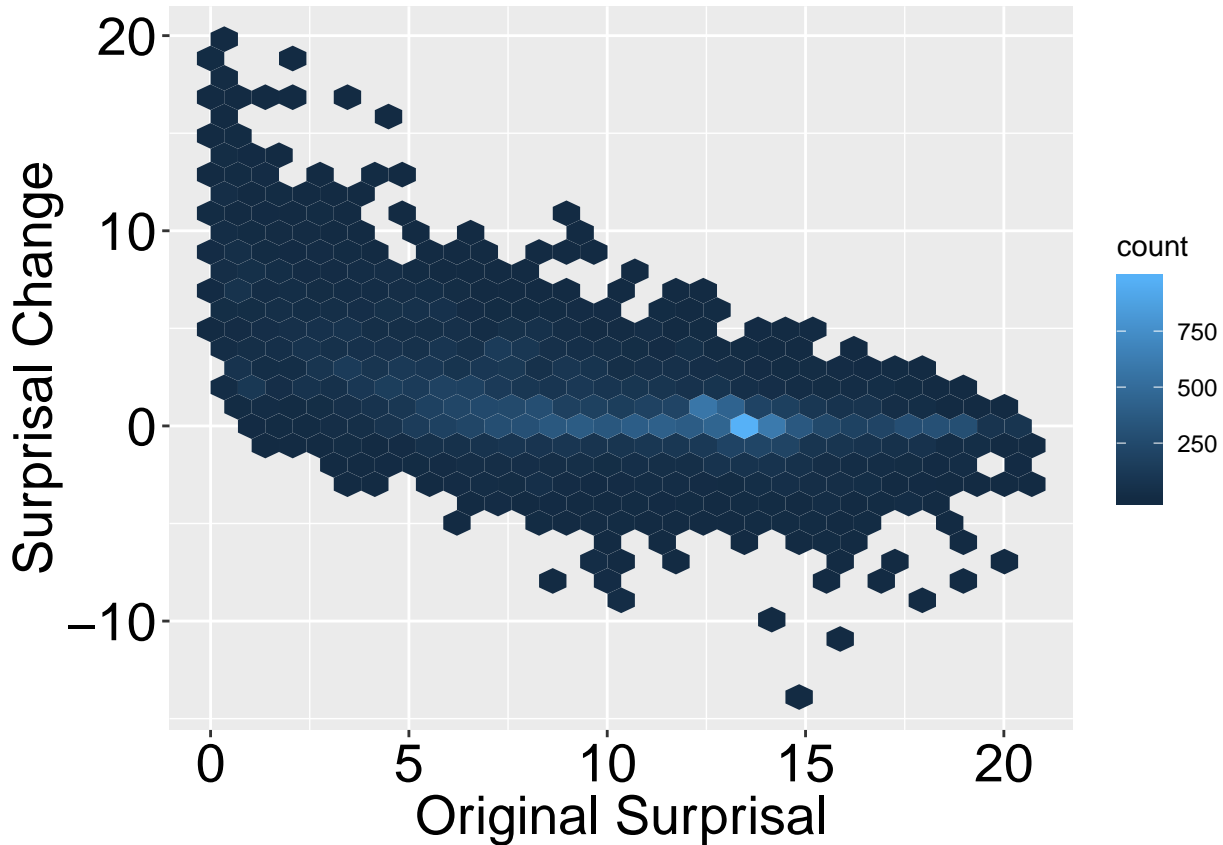
```



```

## t = -61.445, df = 20828, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99.80769 percent confidence interval:
## -1.097492 -0.991998
## sample estimates:
## mean of the differences
## -1.044745
##
##
## Cohen's d
##
## d estimate: -0.4257499 (small)
## 95 percent confidence interval:
## inf sup
## -0.4451725 -0.4063273
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 48067000, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
## 99.80769 percent confidence interval:
## -Inf -0.7085814
## sample estimates:
## (pseudo)median
## -0.7497939
##
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 48067000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
## 99.80769 percent confidence interval:
## -0.7959682 -0.7055539
## sample estimates:
## (pseudo)median
## -0.7497939
##
##
## Cliff's Delta
##
## delta estimate: -0.1240777 (negligible)
## 95 percent confidence interval:
## inf sup
## -0.1350796 -0.1130453

```

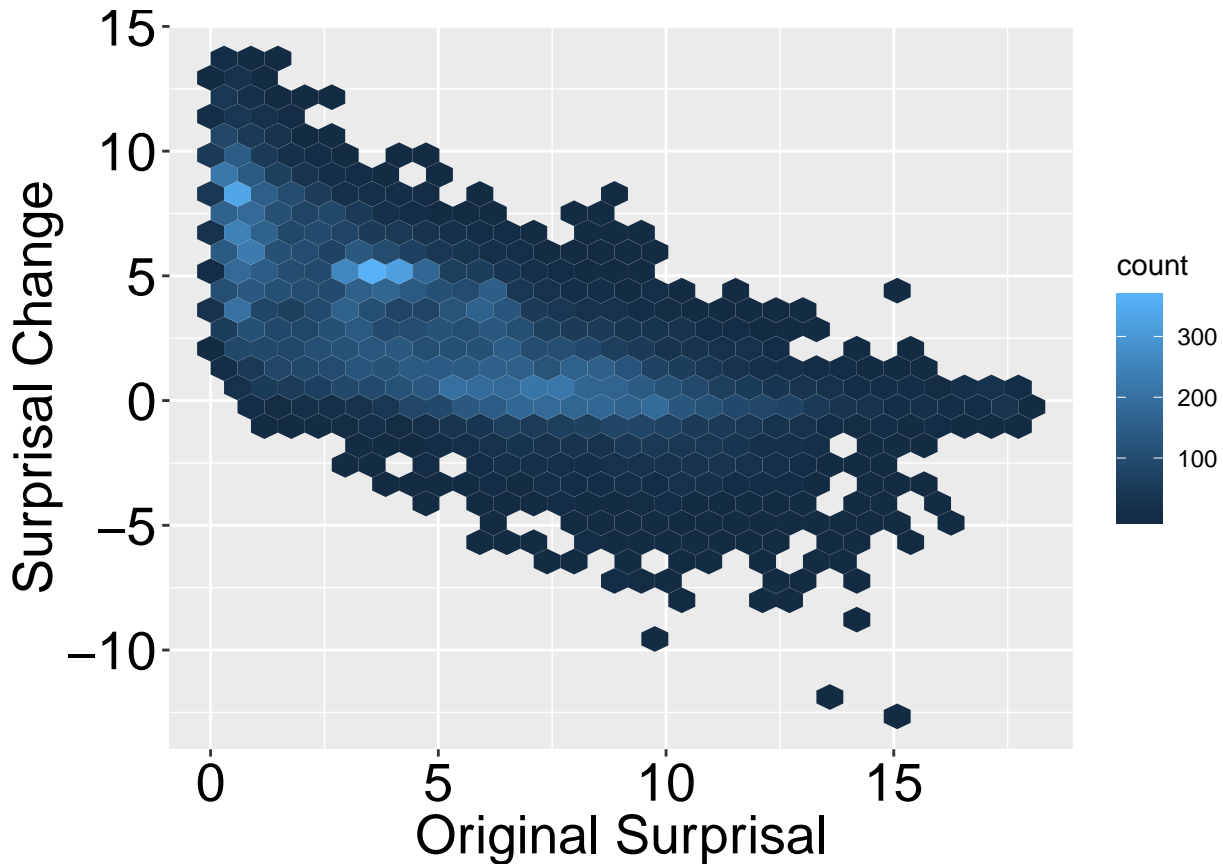


```
## [1] "----- Expression Cache Model -----"
## [1] "ArithSwapTopFiltered100CacheExp Original < Transformed"
##
## Paired t-test
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## t = -136.42, df = 20828, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 99.80769 percent confidence interval:
##      -Inf -2.940254
## sample estimates:
## mean of the differences
##      -3.003908
##
## Paired t-test
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## t = -136.42, df = 20828, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99.80769 percent confidence interval:
## -3.072216 -2.935599
## sample estimates:
## mean of the differences
##      -3.003908
##
##
```

```

## Cohen's d
##
## d estimate: -0.9452715 (large)
## 95 percent confidence interval:
##      inf      sup
## -0.9655219 -0.9250211
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 16014000, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
## 99.80769 percent confidence interval:
##      -Inf -2.83057
## sample estimates:
## (pseudo)median
##      -2.895987
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 16014000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
## 99.80769 percent confidence interval:
##      -2.967289 -2.825819
## sample estimates:
## (pseudo)median
##      -2.895987
##
## Cliff's Delta
##
## delta estimate: -0.5093595 (large)
## 95 percent confidence interval:
##      inf      sup
## -0.5187104 -0.4998868

```

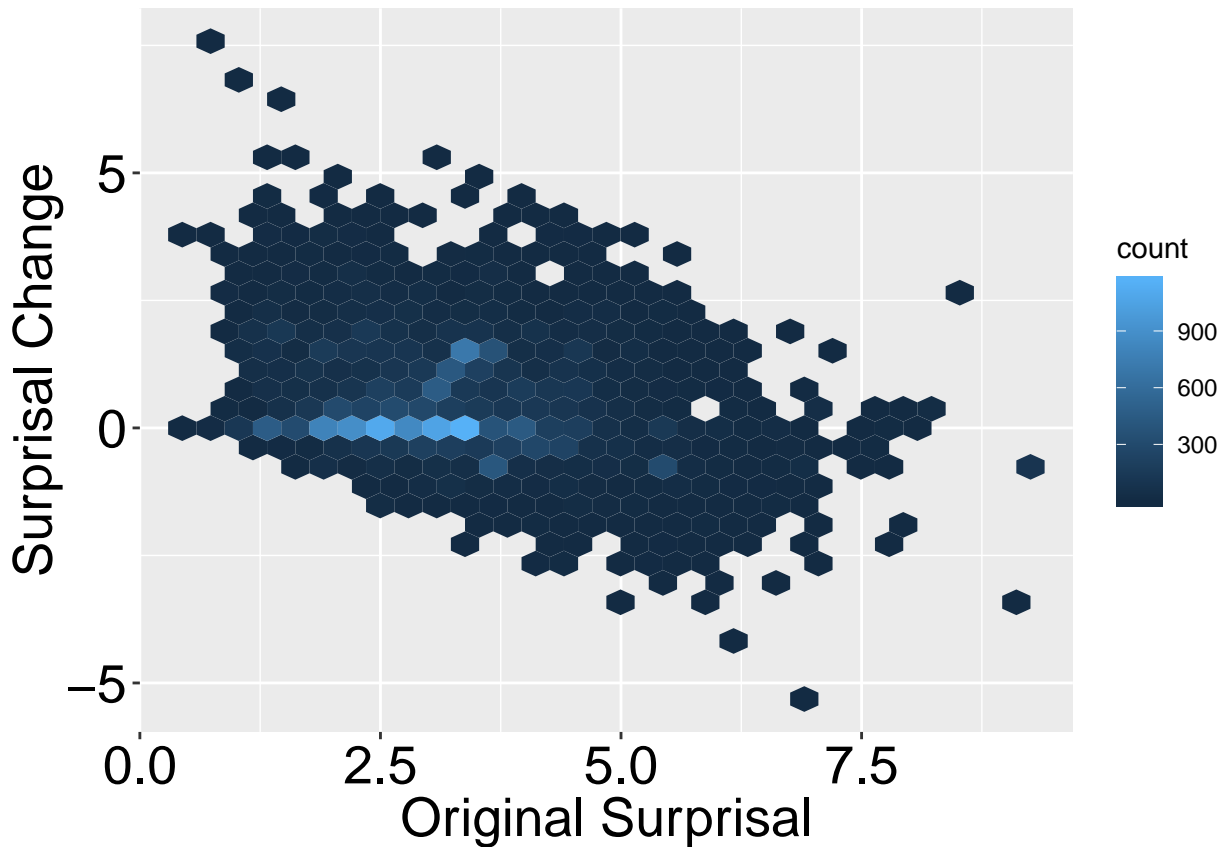


```
## [1] "----- Expression Global Type Model ----- "
## [1] "ArithSwapTopFiltered100GlobalTypeExp Original < Transformed"
##
## Paired t-test
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## t = -67.651, df = 20828, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 99.80769 percent confidence interval:
##      -Inf -0.3952192
## sample estimates:
## mean of the differences
##      -0.4128613
##
## Paired t-test
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## t = -67.651, df = 20828, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99.80769 percent confidence interval:
## -0.4317937 -0.3939290
## sample estimates:
## mean of the differences
##      -0.4128613
##
```

```

##
## Cohen's d
##
## d estimate: -0.4687518 (small)
## 95 percent confidence interval:
##      inf      sup
## -0.4882199 -0.4492836
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 24543000, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
## 99.80769 percent confidence interval:
##      -Inf -0.5079689
## sample estimates:
## (pseudo)median
##      -0.531694
##
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 24543000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
## 99.80769 percent confidence interval:
##      -0.5571657 -0.5060690
## sample estimates:
## (pseudo)median
##      -0.531694
##
##
## Cliff's Delta
##
## delta estimate: -0.1851169 (small)
## 95 percent confidence interval:
##      inf      sup
## -0.1959580 -0.1742306

```

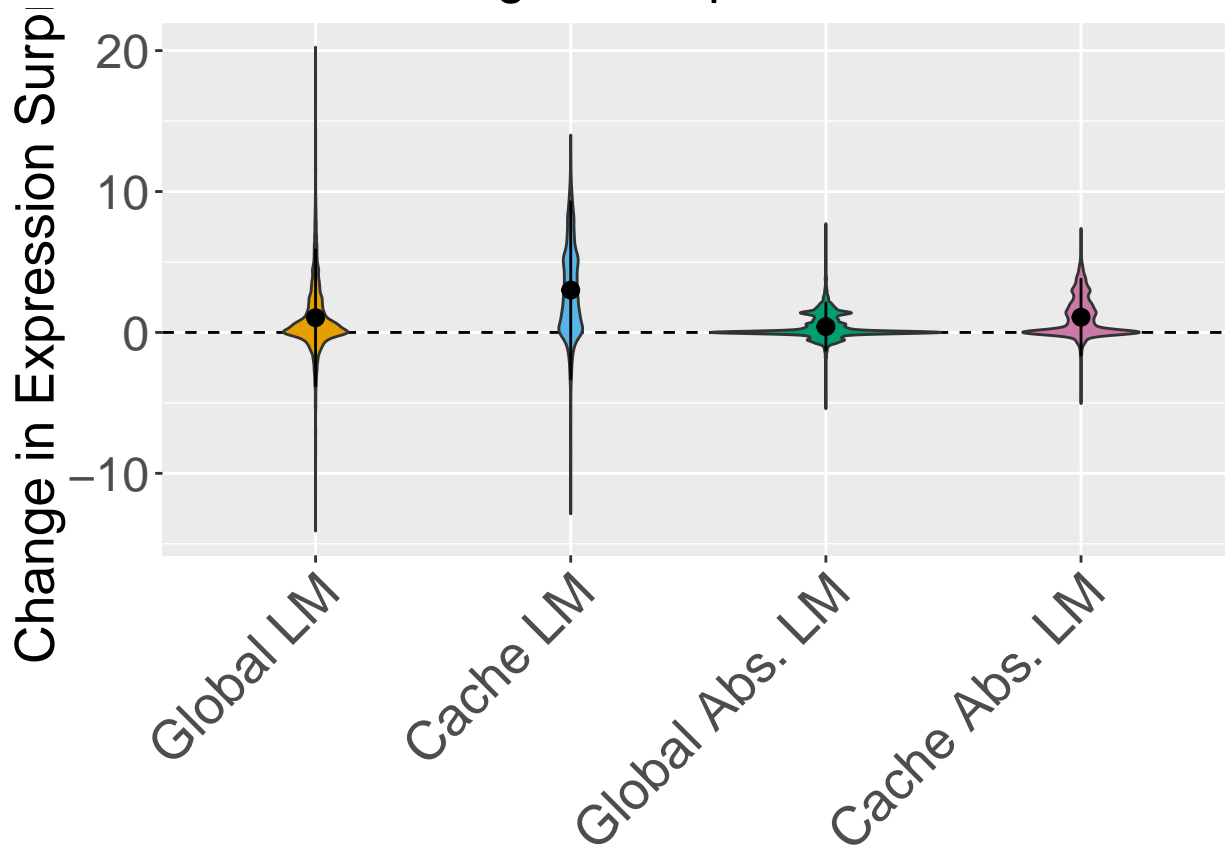
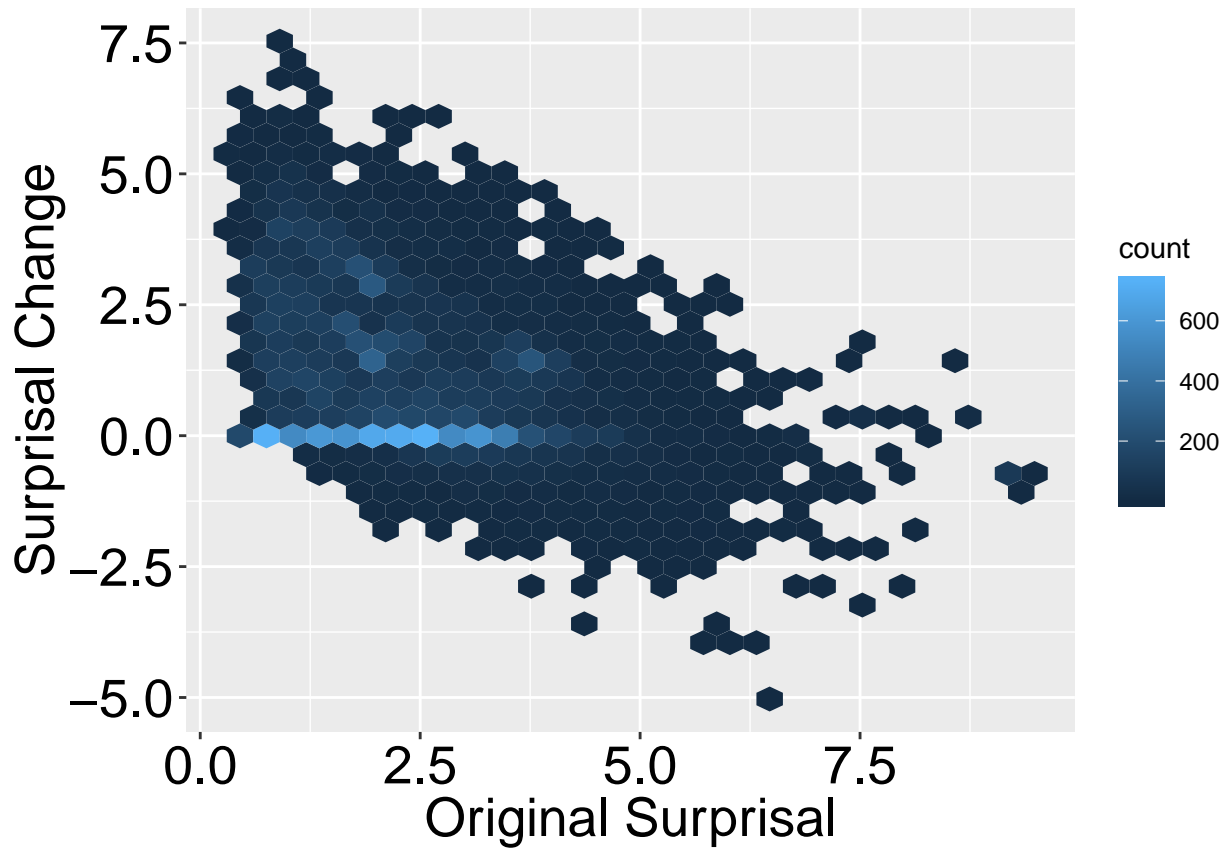


```
## [1] " ----- Expression Cache Type Model ----- "
## [1] "ArithSwapTopFiltered100CacheTypeExp Original < Transformed"
##
## Paired t-test
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## t = -112.55, df = 20828, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 99.80769 percent confidence interval:
##      -Inf -1.053173
## sample estimates:
## mean of the differences
##      -1.080936
##
## Paired t-test
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## t = -112.55, df = 20828, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99.80769 percent confidence interval:
##      -1.110729 -1.051142
## sample estimates:
## mean of the differences
##      -1.080936
##
```

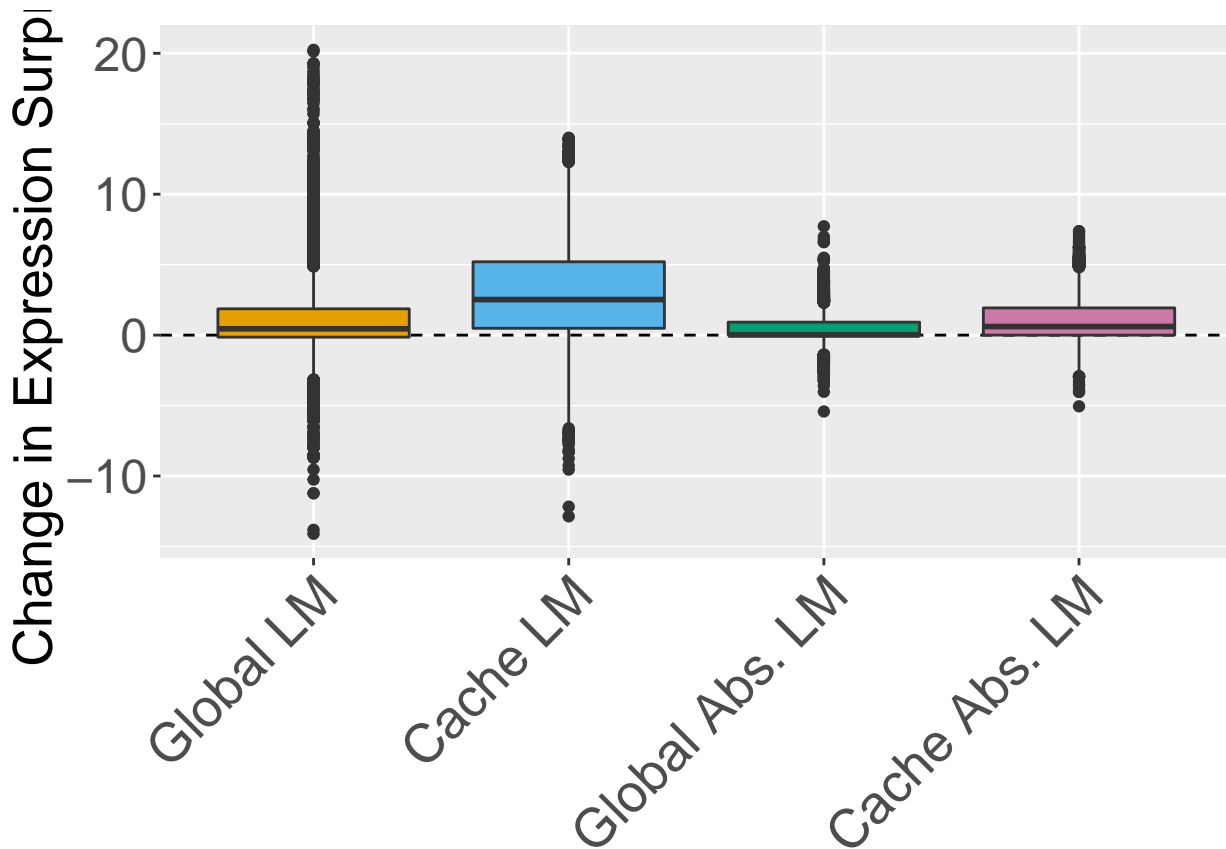
```

## Cohen's d
##
## d estimate: -0.779869 (medium)
## 95 percent confidence interval:
##      inf      sup
## -0.7997919 -0.7599461
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 6654000, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
## 99.80769 percent confidence interval:
##      -Inf -1.393338
## sample estimates:
## (pseudo)median
##      -1.429199
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 6654000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
## 99.80769 percent confidence interval:
##      -1.467706 -1.390706
## sample estimates:
## (pseudo)median
##      -1.429199
##
## Cliff's Delta
##
## delta estimate: -0.428057 (medium)
## 95 percent confidence interval:
##      inf      sup
## -0.4378154 -0.4181977
##
## No id variables; using all as measure variables
## Warning: Ignoring unknown parameters: mult

```







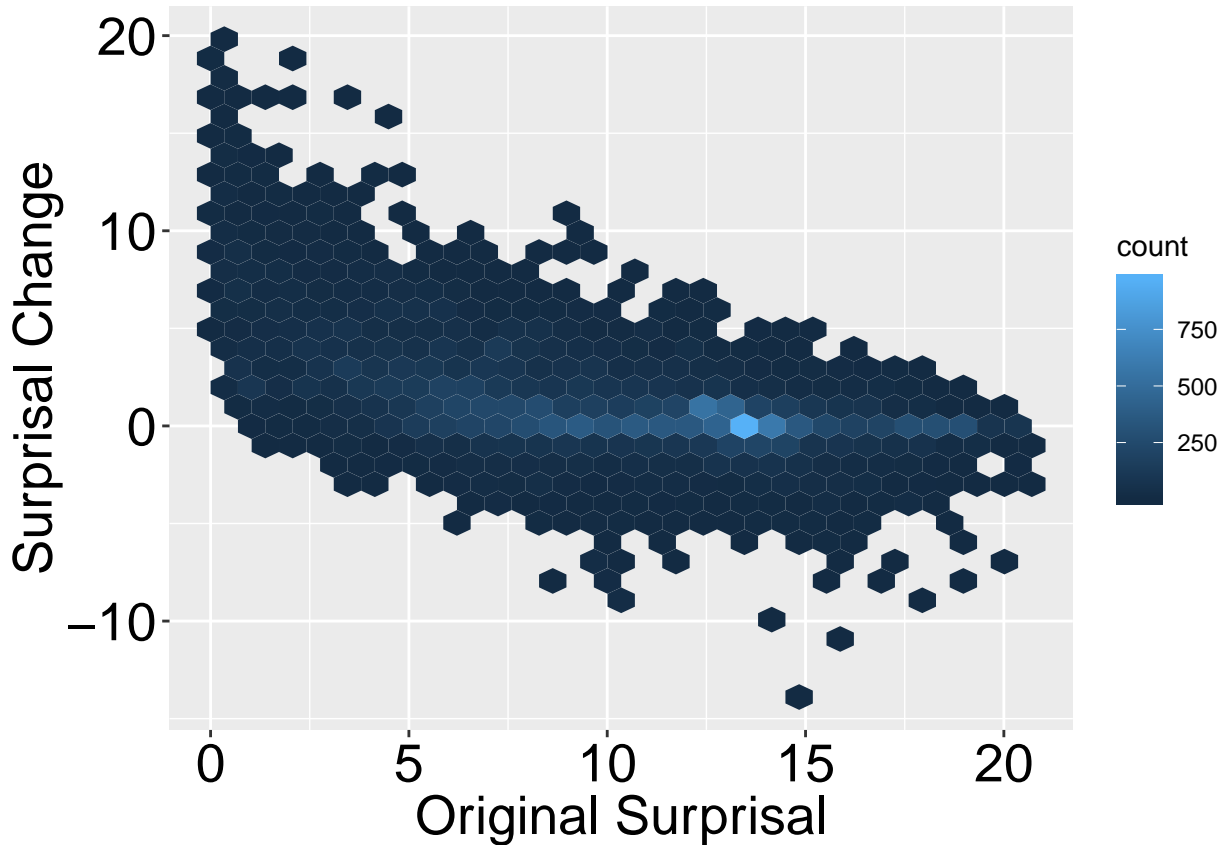
```
dspFiltered2 <- generateFilteredResults(dsp, "Arith", "ArithSwapTopFiltered10", 10)
```

```
## [1] " ----- Expression Global Model ----- "
## [1] "ArithSwapTopFiltered10GlobalExp Original < Transformed"
##
## Paired t-test
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## t = -59.453, df = 19926, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 99.80769 percent confidence interval:
##      -Inf -0.9846969
## sample estimates:
## mean of the differences
##          -1.035024
##
## Paired t-test
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## t = -59.453, df = 19926, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99.80769 percent confidence interval:
## -1.0890311 -0.9810162
## sample estimates:
## mean of the differences
##          -1.035024
```

```

##
##
## Cohen's d
##
## d estimate: -0.4211686 (small)
## 95 percent confidence interval:
##      inf      sup
## -0.4410212 -0.4013160
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 44185000, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
## 99.80769 percent confidence interval:
##      -Inf -0.7022828
## sample estimates:
## (pseudo)median
##      -0.743794
##
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 44185000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
## 99.80769 percent confidence interval:
##      -0.7910414 -0.6993175
## sample estimates:
## (pseudo)median
##      -0.743794
##
##
## Cliff's Delta
##
## delta estimate: -0.1218191 (negligible)
## 95 percent confidence interval:
##      inf      sup
## -0.1330704 -0.1105365

```



```
## [1] "----- Expression Cache Model ----- "
```

```
## [1] "ArithSwapTopFiltered10CacheExp Original < Transformed"
```

```
##
```

```
## Paired t-test
```

```
##
```

```
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
```

```
## t = -131.61, df = 19926, p-value < 2.2e-16
```

```
## alternative hypothesis: true difference in means is less than 0
```

```
## 99.80769 percent confidence interval:
```

```
##      -Inf -2.877827
```

```
## sample estimates:
```

```
## mean of the differences
```

```
##      -2.942459
```

```
##
```

```
##
```

```
## Paired t-test
```

```
##
```

```
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
```

```
## t = -131.61, df = 19926, p-value < 2.2e-16
```

```
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 99.80769 percent confidence interval:
```

```
## -3.011819 -2.873100
```

```
## sample estimates:
```

```
## mean of the differences
```

```
##      -2.942459
```

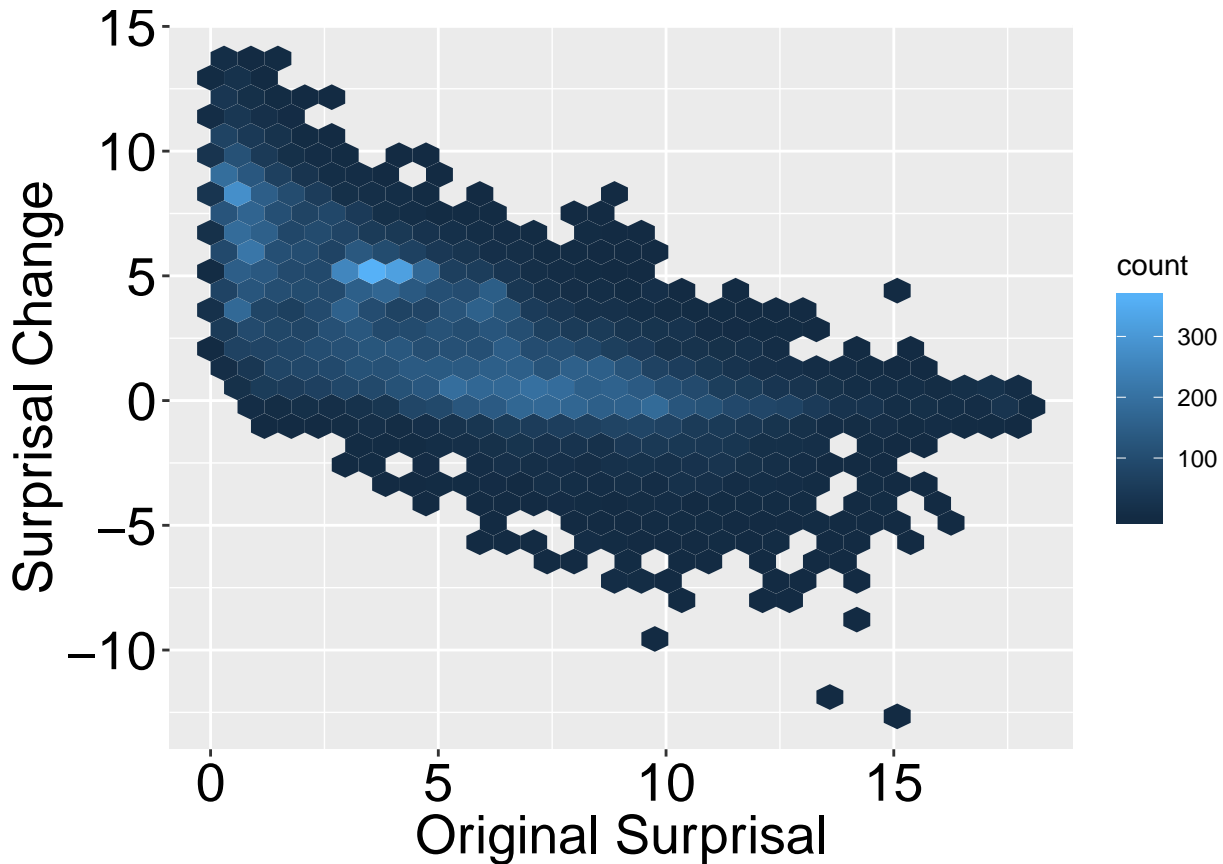
```
##
```

```
##
```

```

## Cohen's d
##
## d estimate: -0.932312 (large)
## 95 percent confidence interval:
##      inf      sup
## -0.9529874 -0.9116367
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 15244000, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
## 99.80769 percent confidence interval:
##      -Inf -2.76919
## sample estimates:
## (pseudo)median
##      -2.834851
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 15244000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
## 99.80769 percent confidence interval:
##      -2.906071 -2.764409
## sample estimates:
## (pseudo)median
##      -2.834851
##
## Cliff's Delta
##
## delta estimate: -0.5029878 (large)
## 95 percent confidence interval:
##      inf      sup
## -0.5125979 -0.4932518

```

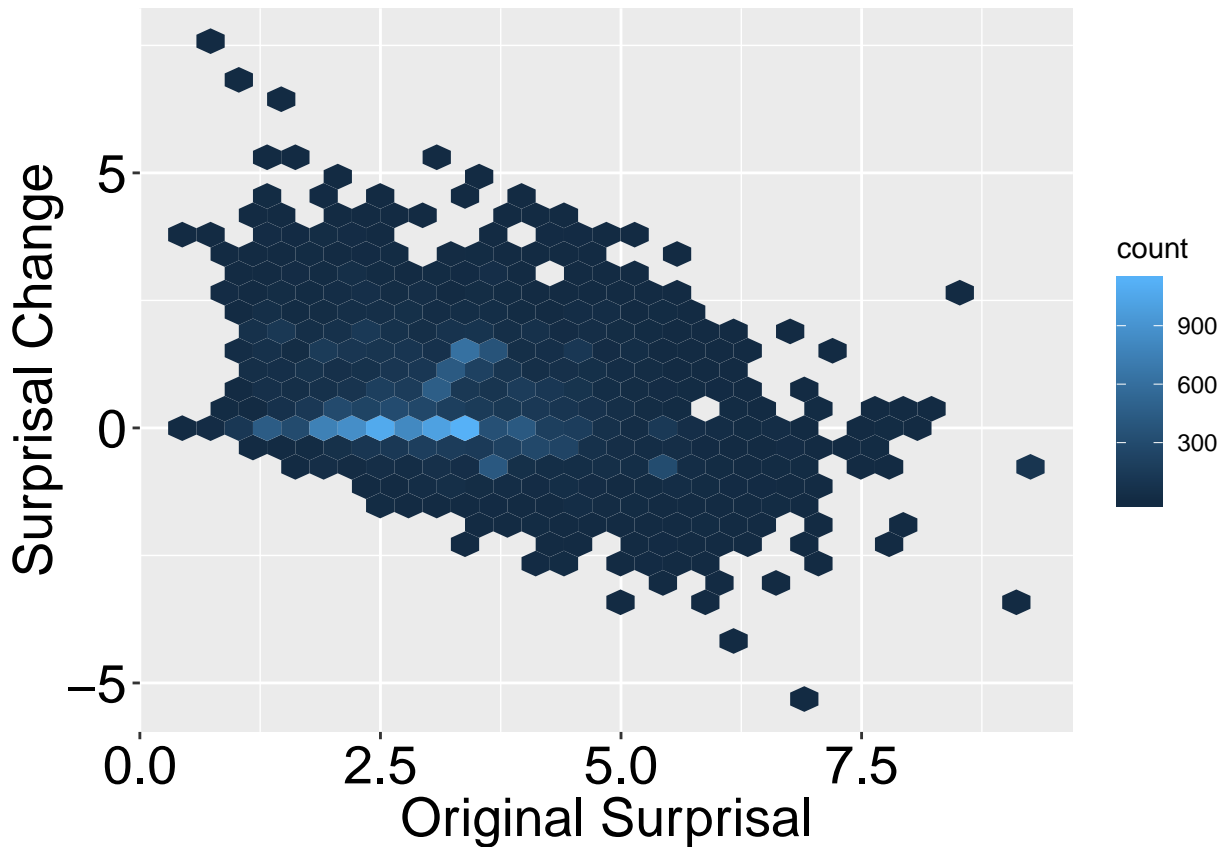


```
## [1] "----- Expression Global Type Model ----- "
## [1] "ArithSwapTopFiltered10GlobalTypeExp Original < Transformed"
##
## Paired t-test
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## t = -66.041, df = 19926, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 99.80769 percent confidence interval:
##      -Inf -0.3959507
## sample estimates:
## mean of the differences
##      -0.4140764
##
## Paired t-test
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## t = -66.041, df = 19926, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99.80769 percent confidence interval:
##      -0.4335277 -0.3946251
## sample estimates:
## mean of the differences
##      -0.4140764
##
```

```

##
## Cohen's d
##
## d estimate: -0.4678336 (small)
## 95 percent confidence interval:
##      inf      sup
## -0.4877365 -0.4479307
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 22037000, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
## 99.80769 percent confidence interval:
##      -Inf -0.5128928
## sample estimates:
## (pseudo)median
##      -0.5372199
##
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 22037000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
## 99.80769 percent confidence interval:
##      -0.5634283 -0.5112588
## sample estimates:
## (pseudo)median
##      -0.5372199
##
##
## Cliff's Delta
##
## delta estimate: -0.1864748 (small)
## 95 percent confidence interval:
##      inf      sup
## -0.1975522 -0.1753497

```



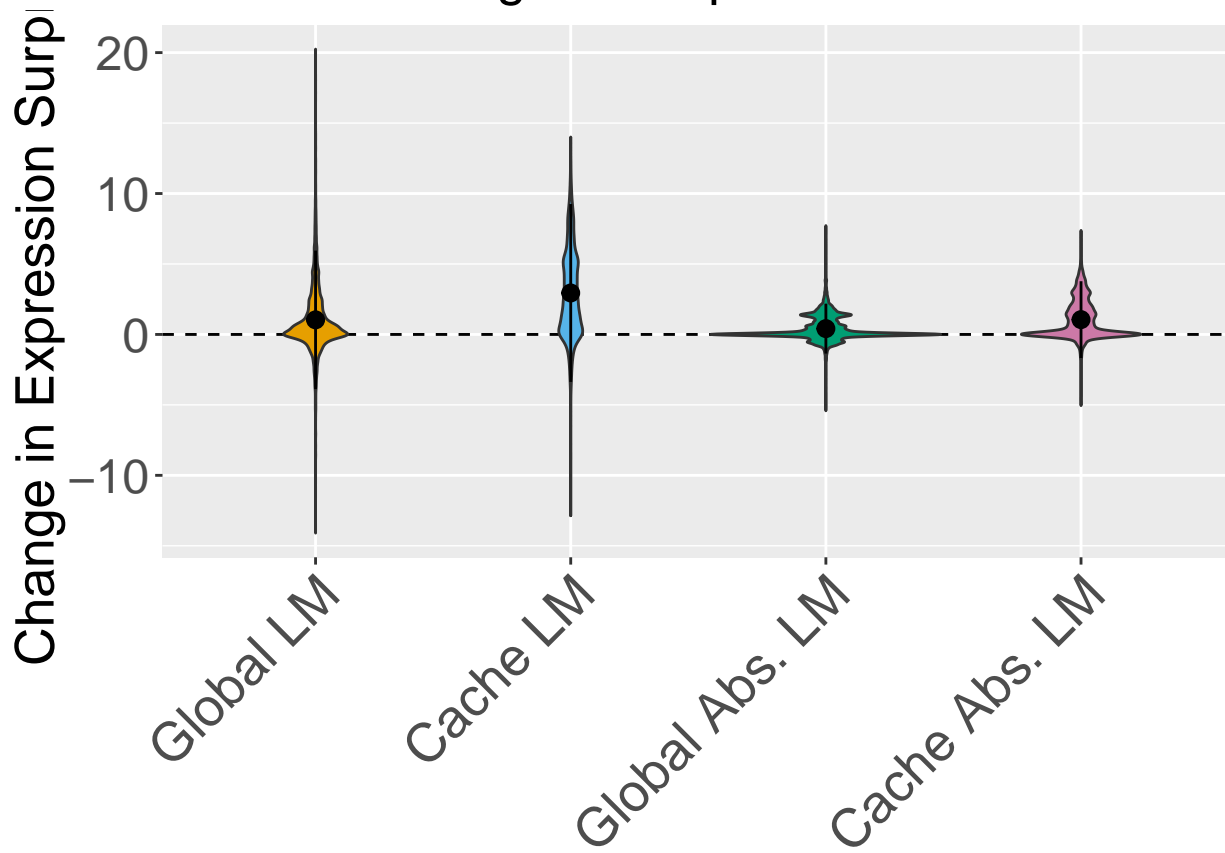
```
## [1] " ----- Expression Cache Type Model ----- "
## [1] "ArithSwapTopFiltered10CacheTypeExp Original < Transformed"
##
## Paired t-test
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## t = -109.03, df = 19926, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 99.80769 percent confidence interval:
##      -Inf -1.023681
## sample estimates:
## mean of the differences
##      -1.051564
##
## Paired t-test
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## t = -109.03, df = 19926, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99.80769 percent confidence interval:
## -1.081486 -1.021642
## sample estimates:
## mean of the differences
##      -1.051564
##
##
```

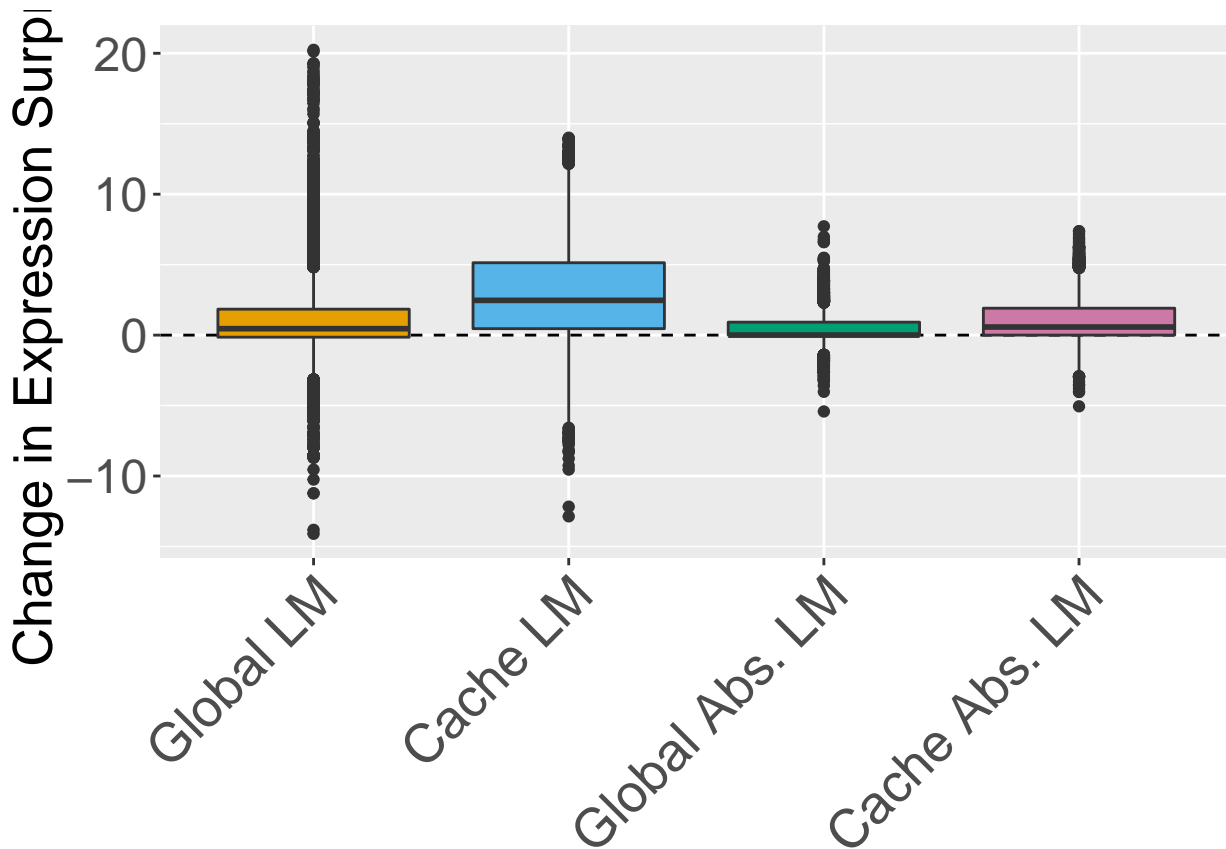
```

## Cohen's d
##
## d estimate: -0.7723378 (medium)
## 95 percent confidence interval:
##      inf      sup
## -0.7926928 -0.7519828
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 6229300, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
## 99.80769 percent confidence interval:
##      -Inf -1.368809
## sample estimates:
## (pseudo)median
##      -1.405054
##
## Wilcoxon signed rank test with continuity correction
##
## data: diffClean$BaseAveEnt and diffClean$ChangeAveEnt
## V = 6229300, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
## 99.80769 percent confidence interval:
##      -1.443762 -1.366163
## sample estimates:
## (pseudo)median
##      -1.405054
##
## Cliff's Delta
##
## delta estimate: -0.4204414 (medium)
## 95 percent confidence interval:
##      inf      sup
## -0.4304698 -0.4103092
##
## No id variables; using all as measure variables
## Warning: Ignoring unknown parameters: mult

```







```
m_swap_no_out <- modelGlobal(dspFiltered, "==", "")
```

```
##
## Call:
## lm(formula = AverageEntChangeExp ~ BaseAveEntExp + log(NumTokens) +
##     factor(ParentOp) + factor(MostFreqOp), data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.4488 -0.8985 -0.0338  0.7990  8.2468
##
## Coefficients:
##
##              Estimate Std. Error t value
## (Intercept)      5.385952   0.184782  29.148
## BaseAveEntExp    -0.278718   0.002741 -101.689
## log(NumTokens)   -0.669708   0.028161 -23.781
## factor(ParentOp)<      0.389141   0.181597   2.143
## factor(ParentOp)<=     0.439838   0.227915   1.930
## factor(ParentOp)>      0.158657   0.195952   0.810
## factor(ParentOp)>=    -0.092584   0.252401  -0.367
## factor(ParentOp)-     0.098817   0.178526   0.554
## factor(ParentOp)/     0.147991   0.204854   0.722
## factor(ParentOp)+    -0.314029   0.171275  -1.833
## factor(ParentOp)ArrayAccess -0.734053   0.167783  -4.375
## factor(ParentOp)ArrayCreation -0.725015   0.186447  -3.889
## factor(ParentOp)Assignment -0.264176   0.167875  -1.574
## factor(ParentOp)ClassInstanceCreation -0.309621   0.188224  -1.645
```

```

## factor(ParentOp)ConditionalExpression      -0.222473   0.203887  -1.091
## factor(ParentOp)MethodInvocation          0.239683   0.163735   1.464
## factor(ParentOp)ParenthesizedExpression  -0.198920   0.166501  -1.195
## factor(ParentOp)ReturnStatement          -1.046165   0.177110  -5.907
## factor(ParentOp)VariableDeclarationFragment -0.587746   0.165543  -3.550
## factor(MostFreqOp)+                       0.636328   0.032258  19.726
##
## Pr(>|t|)
## (Intercept)                               < 2e-16 ***
## BaseAveEntExp                             < 2e-16 ***
## log(NumTokens)                            < 2e-16 ***
## factor(ParentOp)<                          0.032137 *
## factor(ParentOp)<=                         0.053644 .
## factor(ParentOp)>                          0.418140
## factor(ParentOp)>=                         0.713764
## factor(ParentOp)-                          0.579916
## factor(ParentOp)/                          0.470043
## factor(ParentOp)+                          0.066750 .
## factor(ParentOp)ArrayAccess                1.22e-05 ***
## factor(ParentOp)ArrayCreation              0.000101 ***
## factor(ParentOp)Assignment                 0.115590
## factor(ParentOp)ClassInstanceCreation     0.099998 .
## factor(ParentOp)ConditionalExpression      0.275218
## factor(ParentOp)MethodInvocation           0.143256
## factor(ParentOp)ParenthesizedExpression    0.232219
## factor(ParentOp)ReturnStatement            3.55e-09 ***
## factor(ParentOp)VariableDeclarationFragment 0.000386 ***
## factor(MostFreqOp)+                       < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.581 on 16853 degrees of freedom
## Multiple R-squared:  0.4249, Adjusted R-squared:  0.4242
## F-statistic: 655.3 on 19 and 16853 DF, p-value: < 2.2e-16
##
## Analysis of Variance Table
##
## Response: AverageEntChangeExp
##
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## BaseAveEntExp      1  26307  26307.2 10526.419 < 2.2e-16 ***
## log(NumTokens)     1    973    972.9   389.274 < 2.2e-16 ***
## factor(ParentOp)   16   2863   179.0    71.608 < 2.2e-16 ***
## factor(MostFreqOp)  1    972    972.5   389.117 < 2.2e-16 ***
## Residuals         16853  42118     2.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "0.35921999" "0.01328421" "0.03909886" "0.01327884" "0.57511811"
##
##           GVIF Df GVIF^(1/(2*Df))
## BaseAveEntExp      1.082721  1      1.040539
## log(NumTokens)     1.487042  1      1.219443
## factor(ParentOp)   2.114677 16      1.023679
## factor(MostFreqOp) 1.357297  1      1.165031
##
## % Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
## % Date and time: Tue, Feb 19, 2019 - 09:58:28 PM

```

```

## \begin{table}[!htbp] \centering
## \caption{}
## \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lc}
## \[-1.8ex]\hline
## \hline \[-1.8ex]
## & \multicolumn{1}{c}{\textit{Dependent variable:}} \\\
## \cline{2-2}
## \[-1.8ex] & AverageEntChangeExp \\\
## \hline \[-1.8ex]
## BaseAveEntExp &  $-\$0.279^{\***}$  (0.003) \\\
## \log(NumTokens) &  $-\$0.670^{\***}$  (0.028) \\\
## \textless factor(ParentOp) &  $0.389^{\**}$  (0.182) \\\
## \textless factor(ParentOp) &  $= 0.440^{\**}$  (0.228) \\\
## \textgreater factor(ParentOp) &  $0.159$  (0.196) \\\
## \textgreater factor(ParentOp) &  $= -\$0.093$  (0.252) \\\
## factor(ParentOp)- &  $0.099$  (0.179) \\\
## factor(ParentOp)/ &  $0.148$  (0.205) \\\
## factor(ParentOp)+ &  $-\$0.314^{\**}$  (0.171) \\\
## factor(ParentOp)ArrayAccess &  $-\$0.734^{\***}$  (0.168) \\\
## factor(ParentOp)ArrayCreation &  $-\$0.725^{\***}$  (0.186) \\\
## factor(ParentOp)Assignment &  $-\$0.264$  (0.168) \\\
## factor(ParentOp)ClassInstanceCreation &  $-\$0.310^{\**}$  (0.188) \\\
## factor(ParentOp)ConditionalExpression &  $-\$0.222$  (0.204) \\\
## factor(ParentOp)MethodInvocation &  $0.240$  (0.164) \\\
## factor(ParentOp)ParenthesizedExpression &  $-\$0.199$  (0.167) \\\
## factor(ParentOp)ReturnStatement &  $-\$1.046^{\***}$  (0.177) \\\
## factor(ParentOp)VariableDeclarationFragment &  $-\$0.588^{\***}$  (0.166) \\\
## factor(MostFreqOp)+ &  $0.636^{\***}$  (0.032) \\\
## Constant &  $5.386^{\***}$  (0.185) \\\
## \hline \[-1.8ex]
## Observations & 16,873 \\\
## R2 & 0.425 \\\
## Adjusted R2 & 0.424 \\\
## Residual Std. Error & 1.581 (df = 16853) \\\
## F Statistic &  $655.292^{\***}$  (df = 19; 16853) \\\
## \hline
## \hline \[-1.8ex]
## \textit{Note:} & \multicolumn{1}{r}{ $^{\**}p < \$0.1$ ;  $^{\***}p < \$0.05$ ;  $^{\****}p < \$0.01$ } \\\
## \end{tabular}
## \end{table}
## % latex table generated in R 3.4.4 by xtable 1.8-3 package
## % Tue Feb 19 21:58:29 2019
## \begin{table}[ht]
## \centering
## \begin{tabular}{lrrrrr}
## \hline
## & Df & Sum Sq & Mean Sq & F value & Pr(>F) \\\
## \hline
## BaseAveEntExp & 1 & 26307.19 & 26307.19 & 10526.42 & 0.0000 \\\
## \log(NumTokens) & 1 & 972.86 & 972.86 & 389.27 & 0.0000 \\\
## \textless factor(ParentOp) & 16 & 2863.37 & 178.96 & 71.61 & 0.0000 \\\
## \textless factor(MostFreqOp) & 1 & 972.47 & 972.47 & 389.12 & 0.0000 \\\
## \textgreater Residuals & 16853 & 42118.32 & 2.50 & & \\\

```

```

## \hline
## \end{tabular}
## \end{table}
m_swap_cache_no_out <- modelCache(dspFiltered, "==", "")

##
## Call:
## lm(formula = CacheAverageEntChangeExp ~ BaseCacheAveEntExp +
##     log(NumTokens) + factor(ParentOp) + factor(MostFreqOp), data = dataset)
##
## Residuals:
##     Min       1Q   Median       3Q      Max
## -8.3011 -1.2423  0.0012  1.1976  7.0825
##
## Coefficients:
##              Estimate Std. Error t value
## (Intercept)      8.997846   0.246234   36.542
## BaseCacheAveEntExp -0.645297   0.004316 -149.515
## log(NumTokens)   -1.061120   0.034095 -31.123
## factor(ParentOp)<    0.333815   0.245718    1.359
## factor(ParentOp)<=    0.355721   0.306431    1.161
## factor(ParentOp)>   -0.137954   0.261736   -0.527
## factor(ParentOp)>=  -0.213044   0.319907   -0.666
## factor(ParentOp)-    0.243292   0.240962    1.010
## factor(ParentOp)/    0.438962   0.275556    1.593
## factor(ParentOp)+   -0.592341   0.233508   -2.537
## factor(ParentOp)ArrayAccess -0.459026   0.229464   -2.000
## factor(ParentOp)ArrayCreation -1.305937   0.250015   -5.223
## factor(ParentOp)Assignment -1.409897   0.229333   -6.148
## factor(ParentOp)ClassInstanceCreation -0.584215   0.253605   -2.304
## factor(ParentOp)ConditionalExpression -0.116922   0.293150   -0.399
## factor(ParentOp)MethodInvocation    0.353736   0.225253    1.570
## factor(ParentOp)ParenthesizedExpression -0.439583   0.228093   -1.927
## factor(ParentOp)ReturnStatement   -1.892841   0.239497   -7.903
## factor(ParentOp)VariableDeclarationFragment -0.594371   0.227000   -2.618
## factor(MostFreqOp)+    0.765596   0.039180   19.540
##
##              Pr(>|t|)
## (Intercept)      < 2e-16 ***
## BaseCacheAveEntExp < 2e-16 ***
## log(NumTokens)   < 2e-16 ***
## factor(ParentOp)<    0.17431
## factor(ParentOp)<=    0.24572
## factor(ParentOp)>    0.59815
## factor(ParentOp)>=    0.50545
## factor(ParentOp)-    0.31267
## factor(ParentOp)/    0.11118
## factor(ParentOp)+    0.01120 *
## factor(ParentOp)ArrayAccess    0.04547 *
## factor(ParentOp)ArrayCreation 1.78e-07 ***
## factor(ParentOp)Assignment 8.04e-10 ***
## factor(ParentOp)ClassInstanceCreation    0.02126 *
## factor(ParentOp)ConditionalExpression    0.69001
## factor(ParentOp)MethodInvocation    0.11634
## factor(ParentOp)ParenthesizedExpression    0.05397 .

```

```

## factor(ParentOp)ReturnStatement          2.88e-15 ***
## factor(ParentOp)VariableDeclarationFragment 0.00884 **
## factor(MostFreqOp)+                      < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.884 on 16618 degrees of freedom
## Multiple R-squared:  0.6283, Adjusted R-squared:  0.6278
## F-statistic: 1478 on 19 and 16618 DF,  p-value: < 2.2e-16
##
## Analysis of Variance Table
##
## Response: CacheAverageEntChangeExp
##
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## BaseCacheAveEntExp      1  89323   89323 25164.85 < 2.2e-16 ***
## log(NumTokens)          1   2029    2029   571.57 < 2.2e-16 ***
## factor(ParentOp)        16   6985     437   122.98 < 2.2e-16 ***
## factor(MostFreqOp)       1   1355    1355   381.82 < 2.2e-16 ***
## Residuals              16618  58986         4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "0.562921675" "0.012785653" "0.044017440" "0.008541159" "0.371734073"
##           GVIF Df GVIF^(1/(2*Df))
## BaseCacheAveEntExp 1.107542  1      1.052398
## log(NumTokens)    1.493668  1      1.222157
## factor(ParentOp)  2.103933 16      1.023516
## factor(MostFreqOp) 1.375908  1      1.172991
##
## % Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
## % Date and time: Tue, Feb 19, 2019 - 09:58:36 PM
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
##   \begin{tabular}{@{\extracolsep{5pt}}lc}
##     \hline
##     \hline \hline
##     & \multicolumn{1}{c}{\textit{Dependent variable:}} & \hline
##     \cline{2-2}
##     \hline & CacheAverageEntChangeExp & \hline
##     \hline \hline
##     BaseCacheAveEntExp & $-0.645^{***}$ (0.004) & \hline
##     log(NumTokens) & $-1.061^{***}$ (0.034) & \hline
##     factor(ParentOp)\textless & 0.334 (0.246) & \hline
##     factor(ParentOp)\textless = & 0.356 (0.306) & \hline
##     factor(ParentOp)\textgreater & $-0.138 (0.262) & \hline
##     factor(ParentOp)\textgreater = & $-0.213 (0.320) & \hline
##     factor(ParentOp)- & 0.243 (0.241) & \hline
##     factor(ParentOp)/ & 0.439 (0.276) & \hline
##     factor(ParentOp)+ & $-0.592^{**}$ (0.234) & \hline
##     factor(ParentOp)ArrayAccess & $-0.459^{**}$ (0.229) & \hline
##     factor(ParentOp)ArrayCreation & $-1.306^{***}$ (0.250) & \hline
##     factor(ParentOp)Assignment & $-1.410^{***}$ (0.229) & \hline
##     factor(ParentOp)ClassInstanceCreation & $-0.584^{**}$ (0.254) & \hline
##     factor(ParentOp)ConditionalExpression & $-0.117 (0.293) & \hline

```

```

## factor(ParentOp)MethodInvocation & 0.354 (0.225) \\
## factor(ParentOp)ParenthesizedExpression &  $-\$0.440^{\{*\}}$  (0.228) \\
## factor(ParentOp)ReturnStatement &  $-\$1.893^{\{***\}}$  (0.239) \\
## factor(ParentOp)VariableDeclarationFragment &  $-\$0.594^{\{***\}}$  (0.227) \\
## factor(MostFreqOp)+ &  $0.766^{\{***\}}$  (0.039) \\
## Constant &  $8.998^{\{***\}}$  (0.246) \\
## \hline \\[-1.8ex]
## Observations & 16,638 \\
## R2 & 0.628 \\
## Adjusted R2 & 0.628 \\
## Residual Std. Error & 1.884 (df = 16618) \\
## F Statistic &  $1,478.211^{\{***\}}$  (df = 19; 16618) \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:} & \multicolumn{1}{r}{ $^{\{*\}}$   $p < \$0.1$ ;  $^{\{**\}}$   $p < \$0.05$ ;  $^{\{***\}}$   $p < \$0.01$ } \\
## \end{tabular}
## \end{table}
## % latex table generated in R 3.4.4 by xtable 1.8-3 package
## % Tue Feb 19 21:58:36 2019
## \begin{table}[ht]
## \centering
## \begin{tabular}{lrrrrr}
## \hline
## & Df & Sum Sq & Mean Sq & F value & Pr(>$F) \\
## \hline
## BaseCacheAveEntExp & 1 & 89322.90 & 89322.90 & 25164.85 & 0.0000 \\
## log(NumTokens) & 1 & 2028.79 & 2028.79 & 571.57 & 0.0000 \\
## factor(ParentOp) & 16 & 6984.57 & 436.54 & 122.98 & 0.0000 \\
## factor(MostFreqOp) & 1 & 1355.29 & 1355.29 & 381.82 & 0.0000 \\
## Residuals & 16618 & 58985.76 & 3.55 & & \\
## \hline
## \end{tabular}
## \end{table}

```

```

mt_swap_no_out <- modelGlobalType(dspFiltered, "==", "")

```

```

##
## Call:
## lm(formula = TypeAverageEntChangeExp ~ BaseTypeAveEntExp + log(NumTokens) +
##     factor(ParentOp) + factor(MostFreqOp), data = dataset)
##
## Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3245 -0.4928 -0.1327  0.5355  3.2035
##
## Coefficients:
##
##             Estimate Std. Error t value
## (Intercept)      1.770624   0.103902  17.041
## BaseTypeAveEntExp -0.149813   0.005012 -29.889
## log(NumTokens)   -0.145915   0.012136 -12.023
## factor(ParentOp)< -0.971363   0.102845  -9.445
## factor(ParentOp)<= -1.177014   0.116170 -10.132
## factor(ParentOp)> -1.129160   0.107182 -10.535
## factor(ParentOp)>= -0.896743   0.152093  -5.896
## factor(ParentOp)- -1.122062   0.100880 -11.123

```

```

## factor(ParentOp)/          -0.884445  0.107972  -8.191
## factor(ParentOp)+         -0.836446  0.098641  -8.480
## factor(ParentOp)ArrayAccess -0.316541  0.097742  -3.239
## factor(ParentOp)ArrayCreation -0.868872  0.103869  -8.365
## factor(ParentOp)Assignment -0.797913  0.097659  -8.170
## factor(ParentOp)ClassInstanceCreation -1.005285  0.104344  -9.634
## factor(ParentOp)ConditionalExpression -0.723108  0.117554  -6.151
## factor(ParentOp)MethodInvocation -0.506299  0.096243  -5.261
## factor(ParentOp)ParenthesizedExpression -0.892053  0.097406  -9.158
## factor(ParentOp)ReturnStatement -1.000556  0.100276  -9.978
## factor(ParentOp)VariableDeclarationFragment -1.332878  0.096810 -13.768
## factor(MostFreqOp)+       0.362853  0.013711  26.464
##                               Pr(>|t|)
## (Intercept)                < 2e-16 ***
## BaseTypeAveEntExp           < 2e-16 ***
## log(NumTokens)              < 2e-16 ***
## factor(ParentOp)<           < 2e-16 ***
## factor(ParentOp)<=         < 2e-16 ***
## factor(ParentOp)>         < 2e-16 ***
## factor(ParentOp)>=       3.80e-09 ***
## factor(ParentOp)-         < 2e-16 ***
## factor(ParentOp)/         2.76e-16 ***
## factor(ParentOp)+         < 2e-16 ***
## factor(ParentOp)ArrayAccess 0.0012 **
## factor(ParentOp)ArrayCreation < 2e-16 ***
## factor(ParentOp)Assignment 3.29e-16 ***
## factor(ParentOp)ClassInstanceCreation < 2e-16 ***
## factor(ParentOp)ConditionalExpression 7.86e-10 ***
## factor(ParentOp)MethodInvocation 1.45e-07 ***
## factor(ParentOp)ParenthesizedExpression < 2e-16 ***
## factor(ParentOp)ReturnStatement < 2e-16 ***
## factor(ParentOp)VariableDeclarationFragment < 2e-16 ***
## factor(MostFreqOp)+       < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6771 on 16808 degrees of freedom
## Multiple R-squared:  0.2466, Adjusted R-squared:  0.2457
## F-statistic: 289.5 on 19 and 16808 DF,  p-value: < 2.2e-16
##
## Analysis of Variance Table
##
## Response: TypeAverageEntChangeExp
##              Df Sum Sq Mean Sq F value    Pr(>F)
## BaseTypeAveEntExp      1  280.7   280.72  612.309 < 2.2e-16 ***
## log(NumTokens)         1   30.9    30.86   67.309 2.488e-16 ***
## factor(ParentOp)      16 1889.1   118.07  257.536 < 2.2e-16 ***
## factor(MostFreqOp)     1   321.1   321.08  700.351 < 2.2e-16 ***
## Residuals             16808 7705.8    0.46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "0.027447312" "0.003017165" "0.184708354" "0.031393841" "0.753433327"
##              GVIF Df GVIF^(1/(2*Df))
## BaseTypeAveEntExp 1.267769 1          1.125953

```



```

## log(NumTokens)      1.504139  1      1.226434
## factor(ParentOp)   2.497622 16      1.029017
## factor(MostFreqOp) 1.366172  1      1.168834
##
## % Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
## % Date and time: Tue, Feb 19, 2019 - 09:58:39 PM
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lc}
## \hline
## \hline \hline
## & \multicolumn{1}{c}{\textit{Dependent variable:}} & \\
## \cline{2-2}
## \hline \hline
## BaseTypeAveEntExp &  $-\$0.150^{\{***\}}$  & (0.005) \\
## log(NumTokens) &  $-\$0.146^{\{***\}}$  & (0.012) \\
## factor(ParentOp)\textless &  $-\$0.971^{\{***\}}$  & (0.103) \\
## factor(ParentOp)\textless = &  $-\$1.177^{\{***\}}$  & (0.116) \\
## factor(ParentOp)\textgreater &  $-\$1.129^{\{***\}}$  & (0.107) \\
## factor(ParentOp)\textgreater = &  $-\$0.897^{\{***\}}$  & (0.152) \\
## factor(ParentOp)- &  $-\$1.122^{\{***\}}$  & (0.101) \\
## factor(ParentOp)/ &  $-\$0.884^{\{***\}}$  & (0.108) \\
## factor(ParentOp)+ &  $-\$0.836^{\{***\}}$  & (0.099) \\
## factor(ParentOp)ArrayAccess &  $-\$0.317^{\{***\}}$  & (0.098) \\
## factor(ParentOp)ArrayCreation &  $-\$0.869^{\{***\}}$  & (0.104) \\
## factor(ParentOp)Assignment &  $-\$0.798^{\{***\}}$  & (0.098) \\
## factor(ParentOp)ClassInstanceCreation &  $-\$1.005^{\{***\}}$  & (0.104) \\
## factor(ParentOp)ConditionalExpression &  $-\$0.723^{\{***\}}$  & (0.118) \\
## factor(ParentOp)MethodInvocation &  $-\$0.506^{\{***\}}$  & (0.096) \\
## factor(ParentOp)ParenthesizedExpression &  $-\$0.892^{\{***\}}$  & (0.097) \\
## factor(ParentOp)ReturnStatement &  $-\$1.001^{\{***\}}$  & (0.100) \\
## factor(ParentOp)VariableDeclarationFragment &  $-\$1.333^{\{***\}}$  & (0.097) \\
## factor(MostFreqOp)+ &  $0.363^{\{***\}}$  & (0.014) \\
## Constant &  $1.771^{\{***\}}$  & (0.104) \\
## \hline \hline
## Observations & 16,828 & \\
## R2 & 0.247 & \\
## Adjusted R2 & 0.246 & \\
## Residual Std. Error & 0.677 (df = 16808) & \\
## F Statistic & 289.502{***} (df = 19; 16808) & \\
## \hline
## \hline \hline
## \textit{Note:} & \multicolumn{1}{r}{ $^{\{*\}}$   $p < 0.1$ ;  $^{\{**\}}$   $p < 0.05$ ;  $^{\{***\}}$   $p < 0.01$ } & \\
## \end{tabular}
## \end{table}
## % latex table generated in R 3.4.4 by xtable 1.8-3 package
## % Tue Feb 19 21:58:39 2019
## \begin{table}[ht]
## \centering
## \begin{tabular}{lrrrrr}
## \hline
## & Df & Sum Sq & Mean Sq & F value & Pr(>F) \\

```

```
## \hline
## BaseTypeAveEntExp & 1 & 280.72 & 280.72 & 612.31 & 0.0000 \\
## log(NumTokens) & 1 & 30.86 & 30.86 & 67.31 & 0.0000 \\
## factor(ParentOp) & 16 & 1889.12 & 118.07 & 257.54 & 0.0000 \\
## factor(MostFreqOp) & 1 & 321.08 & 321.08 & 700.35 & 0.0000 \\
## Residuals & 16808 & 7705.80 & 0.46 & & \\
## \hline
## \end{tabular}
## \end{table}
```

```
mt_swap_cache_no_out <- modelCacheType(dspFiltered, "==", "")
```

```
##
## Call:
## lm(formula = CacheTypeAverageEntChangeExp ~ BaseCacheTypeAveEntExp +
## log(NumTokens) + factor(ParentOp) + factor(MostFreqOp), data = dataset)
##
## Residuals:
## Min 1Q Median 3Q Max
## -3.0152 -0.8564 -0.0519 0.7900 4.1332
##
## Coefficients:
## Estimate Std. Error t value
## (Intercept) 3.435967 0.178813 19.215
## BaseCacheTypeAveEntExp -0.401978 0.007482 -53.728
## log(NumTokens) -0.464523 0.020966 -22.156
## factor(ParentOp)< -1.100751 0.178793 -6.157
## factor(ParentOp)<= -1.348488 0.212823 -6.336
## factor(ParentOp)> -1.399130 0.186699 -7.494
## factor(ParentOp)>= -1.011379 0.230705 -4.384
## factor(ParentOp)- -1.316671 0.175019 -7.523
## factor(ParentOp)/ -0.717602 0.187408 -3.829
## factor(ParentOp)+ -0.524371 0.171274 -3.062
## factor(ParentOp)ArrayAccess -0.267172 0.169395 -1.577
## factor(ParentOp)ArrayCreation -1.203929 0.180026 -6.688
## factor(ParentOp)Assignment -1.085298 0.169315 -6.410
## factor(ParentOp)ClassInstanceCreation -0.834076 0.181022 -4.608
## factor(ParentOp)ConditionalExpression -0.594174 0.199530 -2.978
## factor(ParentOp)MethodInvocation -0.178374 0.166955 -1.068
## factor(ParentOp)ParenthesizedExpression -0.937796 0.168768 -5.557
## factor(ParentOp)ReturnStatement -1.325867 0.174226 -7.610
## factor(ParentOp)VariableDeclarationFragment -0.859338 0.168150 -5.111
## factor(MostFreqOp)+ 0.608279 0.023802 25.556
## Pr(>|t|)
## (Intercept) < 2e-16 ***
## BaseCacheTypeAveEntExp < 2e-16 ***
## log(NumTokens) < 2e-16 ***
## factor(ParentOp)< 7.60e-10 ***
## factor(ParentOp)<= 2.41e-10 ***
## factor(ParentOp)> 7.01e-14 ***
## factor(ParentOp)>= 1.17e-05 ***
## factor(ParentOp)- 5.62e-14 ***
## factor(ParentOp)/ 0.000129 ***
## factor(ParentOp)+ 0.002205 **
## factor(ParentOp)ArrayAccess 0.114766
```

```

## factor(ParentOp)ArrayCreation          2.34e-11 ***
## factor(ParentOp)Assignment             1.49e-10 ***
## factor(ParentOp)ClassInstanceCreation  4.10e-06 ***
## factor(ParentOp)ConditionalExpression   0.002907 **
## factor(ParentOp)MethodInvocation        0.285356
## factor(ParentOp)ParenthesizedExpression 2.79e-08 ***
## factor(ParentOp)ReturnStatement         2.88e-14 ***
## factor(ParentOp)VariableDeclarationFragment 3.25e-07 ***
## factor(MostFreqOp)+                     < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.174 on 16900 degrees of freedom
## Multiple R-squared:  0.2672, Adjusted R-squared:  0.2663
## F-statistic: 324.3 on 19 and 16900 DF,  p-value: < 2.2e-16
##
## Analysis of Variance Table
##
## Response: CacheTypeAverageEntChangeExp
##
##           Df Sum Sq Mean Sq F value    Pr(>F)
## BaseCacheTypeAveEntExp      1  3754.5   3754.5  2723.84 < 2.2e-16 ***
## log(NumTokens)              1   325.8    325.8   236.34 < 2.2e-16 ***
## factor(ParentOp)            16  3511.4    219.5   159.22 < 2.2e-16 ***
## factor(MostFreqOp)          1   900.2    900.2   653.09 < 2.2e-16 ***
## Residuals                   16900 23294.6     1.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "0.11811570" "0.01024855" "0.11046851" "0.02832056" "0.73284668"
##
##           GVIF Df GVIF^(1/(2*Df))
## BaseCacheTypeAveEntExp 1.127536  1     1.061855
## log(NumTokens)         1.487928  1     1.219806
## factor(ParentOp)       2.259340 16     1.025798
## factor(MostFreqOp)     1.369888  1     1.170422
##
## % Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
## % Date and time: Tue, Feb 19, 2019 - 09:58:41 PM
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
##   \begin{tabular}{@{\extracolsep{5pt}}lc}
##     \hline
##     & \multicolumn{1}{c}{\textit{Dependent variable:}} & \\
##     \hline
##     \hline
##     & CacheTypeAverageEntChangeExp & \\
##     \hline
##     BaseCacheTypeAveEntExp &  $-\$0.402^{\{***\}}$  & (0.007) \\
##     log(NumTokens) &  $-\$0.465^{\{***\}}$  & (0.021) \\
##     factor(ParentOp)\textless &  $-\$1.101^{\{***\}}$  & (0.179) \\
##     factor(ParentOp)\textless = &  $-\$1.348^{\{***\}}$  & (0.213) \\
##     factor(ParentOp)\textgreater &  $-\$1.399^{\{***\}}$  & (0.187) \\
##     factor(ParentOp)\textgreater = &  $-\$1.011^{\{***\}}$  & (0.231) \\
##     factor(ParentOp)- &  $-\$1.317^{\{***\}}$  & (0.175) \\
##     factor(ParentOp)/ &  $-\$0.718^{\{***\}}$  & (0.187)

```

```

## factor(ParentOp)+ & $-0.524$^{***}$ (0.171) \\
## factor(ParentOp)ArrayAccess & $-0.267 (0.169) \\
## factor(ParentOp)ArrayCreation & $-1.204$^{***}$ (0.180) \\
## factor(ParentOp)Assignment & $-1.085$^{***}$ (0.169) \\
## factor(ParentOp)ClassInstanceCreation & $-0.834$^{***}$ (0.181) \\
## factor(ParentOp)ConditionalExpression & $-0.594$^{***}$ (0.200) \\
## factor(ParentOp)MethodInvocation & $-0.178 (0.167) \\
## factor(ParentOp)ParenthesizedExpression & $-0.938$^{***}$ (0.169) \\
## factor(ParentOp)ReturnStatement & $-1.326$^{***}$ (0.174) \\
## factor(ParentOp)VariableDeclarationFragment & $-0.859$^{***}$ (0.168) \\
## factor(MostFreqOp)+ & 0.608$^{***}$ (0.024) \\
## Constant & 3.436$^{***}$ (0.179) \\
## \hline \\[-1.8ex]
## Observations & 16,920 \\
## R$^{2}$ & 0.267 \\
## Adjusted R$^{2}$ & 0.266 \\
## Residual Std. Error & 1.174 (df = 16900) \\
## F Statistic & 324.250$^{***}$ (df = 19; 16900) \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:} & \multicolumn{1}{r}{\textit{\$}^{*}\textit{\$}p<\$0.1; \textit{\$}^{*}\textit{\$}p<\$0.05; \textit{\$}^{*}\textit{\$}p<\$0.01} \\
## \end{tabular}
## \end{table}
## % latex table generated in R 3.4.4 by xtable 1.8-3 package
## % Tue Feb 19 21:58:41 2019
## \begin{table}[ht]
## \centering
## \begin{tabular}{lrrrrr}
## \hline
## & Df & Sum Sq & Mean Sq & F value & Pr(>F) \\
## \hline
## BaseCacheTypeAveEntExp & 1 & 3754.48 & 3754.48 & 2723.84 & 0.0000 \\
## log(NumTokens) & 1 & 325.77 & 325.77 & 236.34 & 0.0000 \\
## factor(ParentOp) & 16 & 3511.40 & 219.46 & 159.22 & 0.0000 \\
## factor(MostFreqOp) & 1 & 900.21 & 900.21 & 653.09 & 0.0000 \\
## Residuals & 16900 & 23294.61 & 1.38 & & \\
## \hline
## \end{tabular}
## \end{table}
arithOut <- printEffTable(pairedResults)

## [1] "Type,PTOne,PTTwo,CITTwo,CohensD,PWilcoxOne,PWilcoxTwo,CIWilcoxTwo,CliffDelta"
## [1] "ArithSwapTopCacheExp,0,0,-3.3873 -3.3294,-1.3927,0,0,-3.5332 -3.4927,-0.6033"
## [1] "ArithSwapTopCacheTypeExp,0,0,-0.4979 -0.4747,-0.5052,0,0,-1.2551 -1.2173,-0.1263"
## [1] "ArithSwapTopFiltered100CacheExp,0,0,-3.0722 -2.9356,-0.9453,0,0,-2.9673 -2.8258,-0.5094"
## [1] "ArithSwapTopFiltered100CacheTypeExp,0,0,-1.1107 -1.0511,-0.7799,0,0,-1.4677 -1.3907,-0.4281"
## [1] "ArithSwapTopFiltered100GlobalExp,0,0,-1.0975 -0.992,-0.4257,0,0,-0.796 -0.7056,-0.1241"
## [1] "ArithSwapTopFiltered100GlobalTypeExp,0,0,-0.4318 -0.3939,-0.4688,0,0,-0.5572 -0.5061,-0.1851"
## [1] "ArithSwapTopFiltered10CacheExp,0,0,-3.0118 -2.8731,-0.9323,0,0,-2.9061 -2.7644,-0.503"
## [1] "ArithSwapTopFiltered10CacheTypeExp,0,0,-1.0815 -1.0216,-0.7723,0,0,-1.4438 -1.3662,-0.4204"
## [1] "ArithSwapTopFiltered10GlobalExp,0,0,-1.089 -0.981,-0.4212,0,0,-0.791 -0.6993,-0.1218"
## [1] "ArithSwapTopFiltered10GlobalTypeExp,0,0,-0.4335 -0.3946,-0.4678,0,0,-0.5634 -0.5113,-0.1865"
## [1] "ArithSwapTopGlobalExp,0,0,-3.2244 -3.1637,-1.2638,0,0,-3.1225 -3.1226,-0.4337"
## [1] "ArithSwapTopGlobalTypeExp,0,0,-0.2628 -0.2482,-0.4195,0,0,-0.7712 -0.7299,-0.0979"

```

```
save(arithOut, file = "/data/anon/SemanticTransformation/sample/arithOut.RDat")
save(dsp, file = "/data/anon/SemanticTransformation/sample/dsp.RDat")
save(dspFiltered, file = "/data/anon/SemanticTransformation/sample/dspFiltered.RDat")
save(dspFiltered2, file = "/data/anon/SemanticTransformation/sample/dspFiltered2.RDat")
```