BangAndOlufsen

Henry Reith

November 20, 2018

Cleaning Data

Got rid of unconsistencies, NA's, and items that we did not feel were useful. Turned installs from a numeric to factor. Added new column that put the number of updates into tiers and removed things that had a much higher of installs than others.

```
## Warning: package 'caret' was built under R version 3.5.1
```

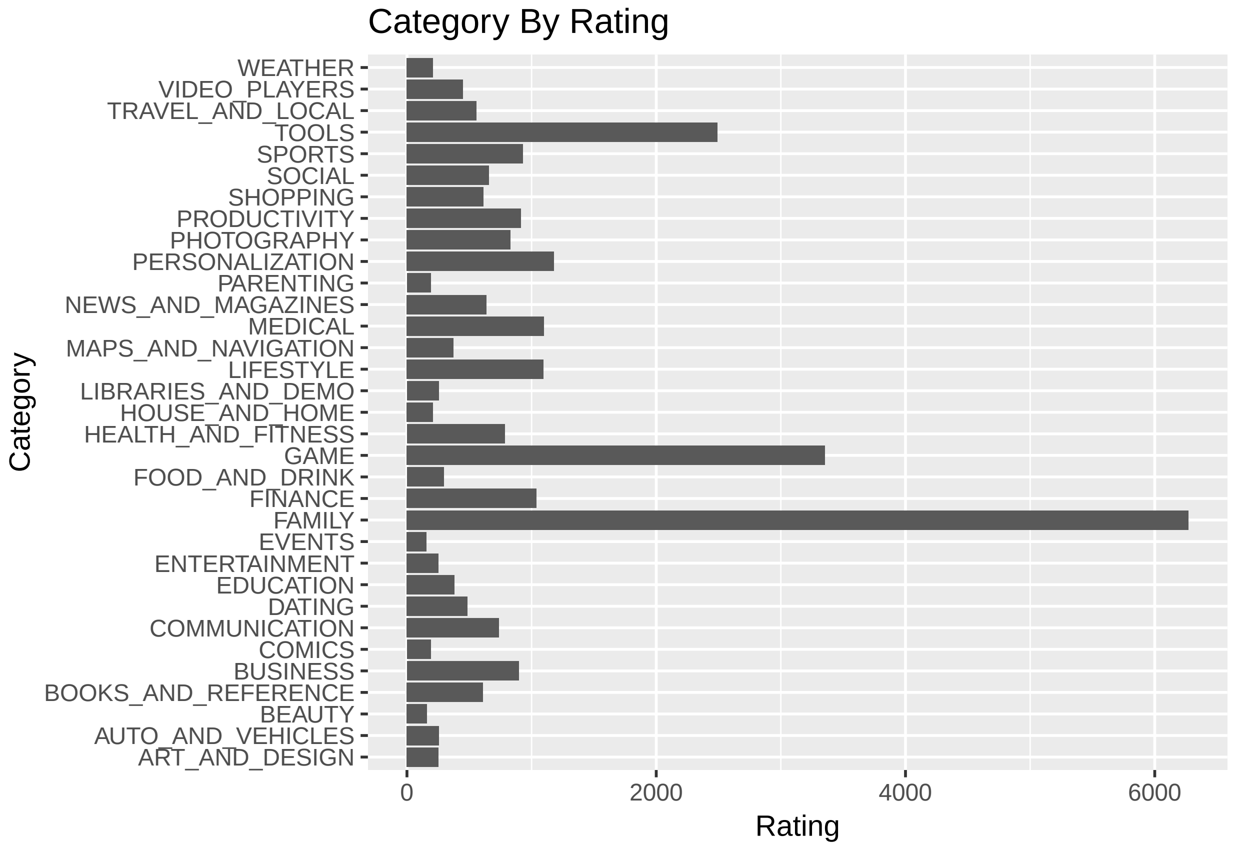
```
## Loading required package: lattice
```

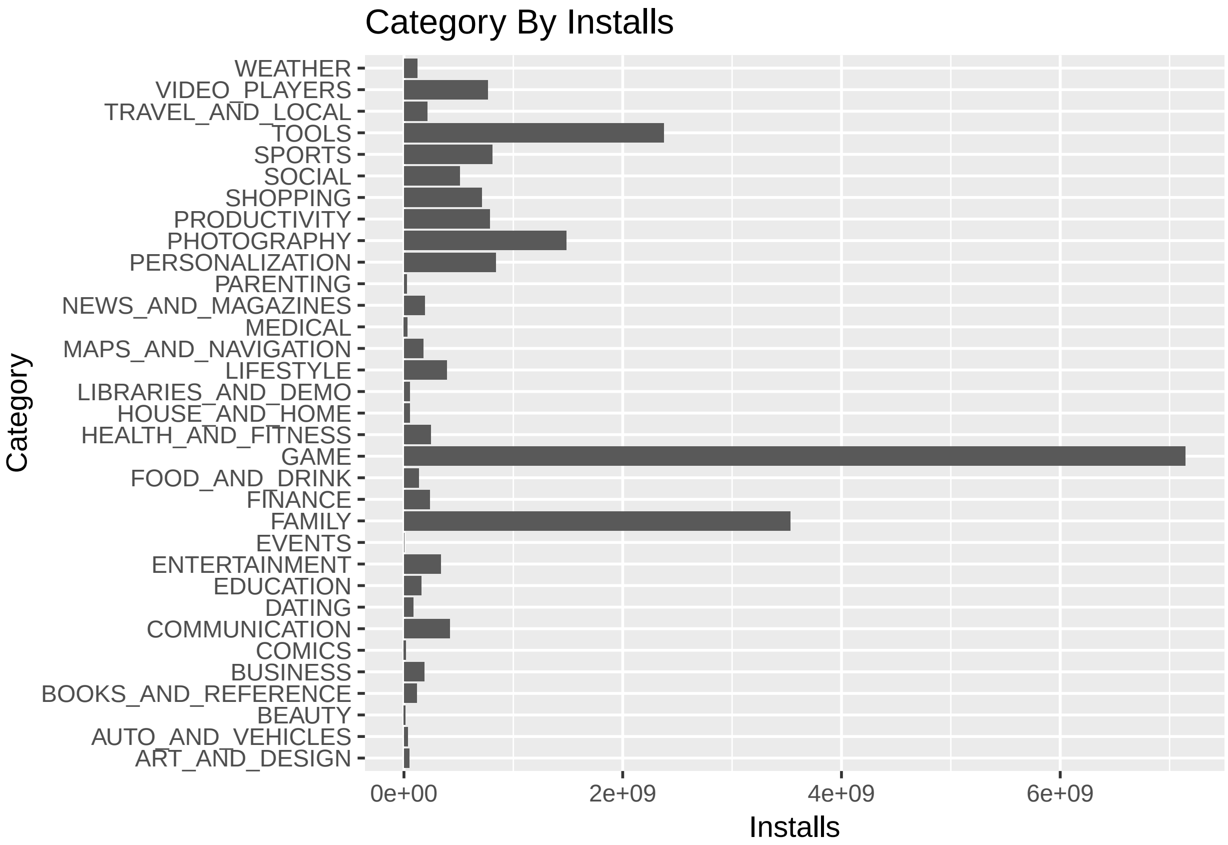
```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.5.1
```

Bar Plots

```
## Warning: package 'ggthemes' was built under R version 3.5.1
```





Random Forest Classifier

Created a random forest classifier to predict which features determined Installs. We were succesful in building the model but could not generate a confusion matrix. Attempts for confusion matrix are commented out in code. This was a successful model with a 75% Variables explained.

```
## Warning: package 'randomForest' was built under R version 3.5.1
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
##
## Call:
##  randomForest(formula = Installs ~ Category + Type + Rating +      Reviews, data = TrainSet, nTrees =
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 3.385096e+13
##                     % Var explained: 75.55
```

RPart Model Before Removal of Outliers

Created RPart model with RSquared value of .49

```
## CART
##
## 6921 samples
## 1918 predictors
##
## No pre-processing
## Resampling: Cross-Validated (4 fold)
## Summary of sample sizes: 5190, 5191, 5191, 5191
## Resampling results across tuning parameters:
##
##   cp            RMSE     Rsquared   MAE
##   5.986376e-12  6345863  0.7086389  1617284
##   1.820065e-11  6345863  0.7086389  1617276
##   6.579333e-10  6345863  0.7086389  1617299
##   1.645620e-09  6345863  0.7086389  1617249
##   2.262403e-09  6345863  0.7086389  1617278
##   1.826979e-07  6345838  0.7086422  1616771
##   2.915690e-07  6345935  0.7086319  1616822
##   5.921409e-06  6346357  0.7085860  1617063
##   1.891030e-04  6333252  0.7095951  1666618
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was cp = 0.000189103.
```

RPart Unsuccessful

Created RPart model with RSquared Value of .07. This model did not include our category of Reviews which shows the importance of this variable.

```
## CART
##
## 6921 samples
##   34 predictor
##
## No pre-processing
## Resampling: Cross-Validated (4 fold)
## Summary of sample sizes: 5192, 5191, 5191, 5189
## Resampling results across tuning parameters:
##
##   cp            RMSE      Rsquared    MAE
##   1.732176e-07  11293916  0.07318876  4336603
##   2.793866e-07  11293823  0.07320068  4336461
##   8.994665e-07  11293852  0.07319531  4336929
##   1.060652e-06  11294356  0.07313268  4336993
##   9.265291e-06  11294597  0.07304075  4335027
##   1.572876e-05  11294109  0.07306509  4334010
##   2.011649e-05  11294067  0.07303052  4332145
##   3.032684e-04  11311964  0.07012274  4357452
##   5.283633e-04  11324262  0.06720517  4379752
##   7.723539e-04  11313896  0.06802394  4380153
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was cp = 2.793866e-07.
```

After Removal of Outliers

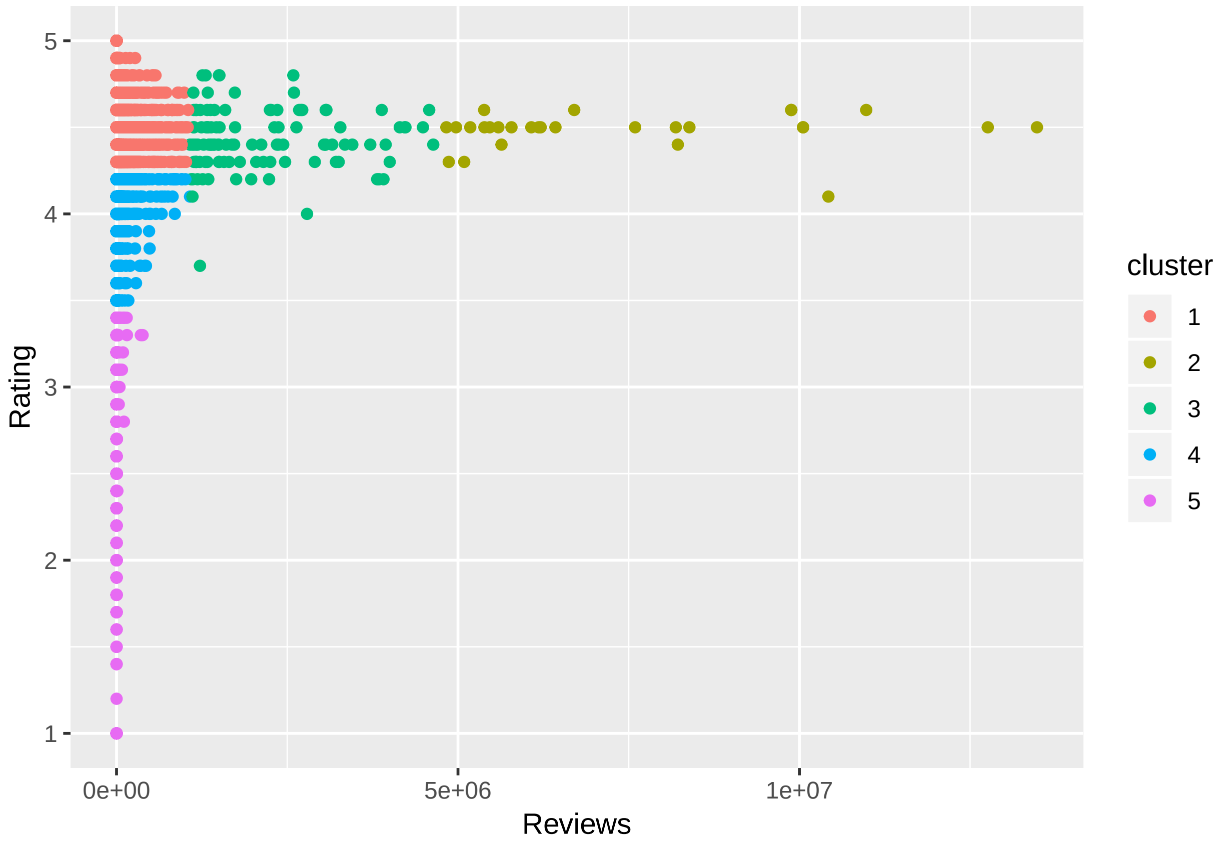
RPart model with RSquared value of .69. This model was created after removing outliers and includes Reviews as a factor.

```
## CART
##
## 6921 samples
##   35 predictor
##
## No pre-processing
## Resampling: Cross-Validated (4 fold)
## Summary of sample sizes: 5190, 5191, 5191, 5191
## Resampling results across tuning parameters:
##
##   cp            RMSE     Rsquared   MAE
##   7.634509e-12  6493429  0.6945728  1705240
##   2.664735e-11  6493429  0.6945728  1705232
##   1.394441e-09  6493428  0.6945732  1705162
##   2.681428e-09  6493428  0.6945733  1705172
##   3.352496e-09  6493427  0.6945733  1705157
##   2.003040e-07  6493394  0.6945769  1703739
##   3.293002e-07  6493338  0.6945825  1703787
##   7.707849e-06  6491577  0.6947318  1705642
##   2.557046e-04  6470648  0.6963449  1737011
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was cp = 0.0002557046.
```

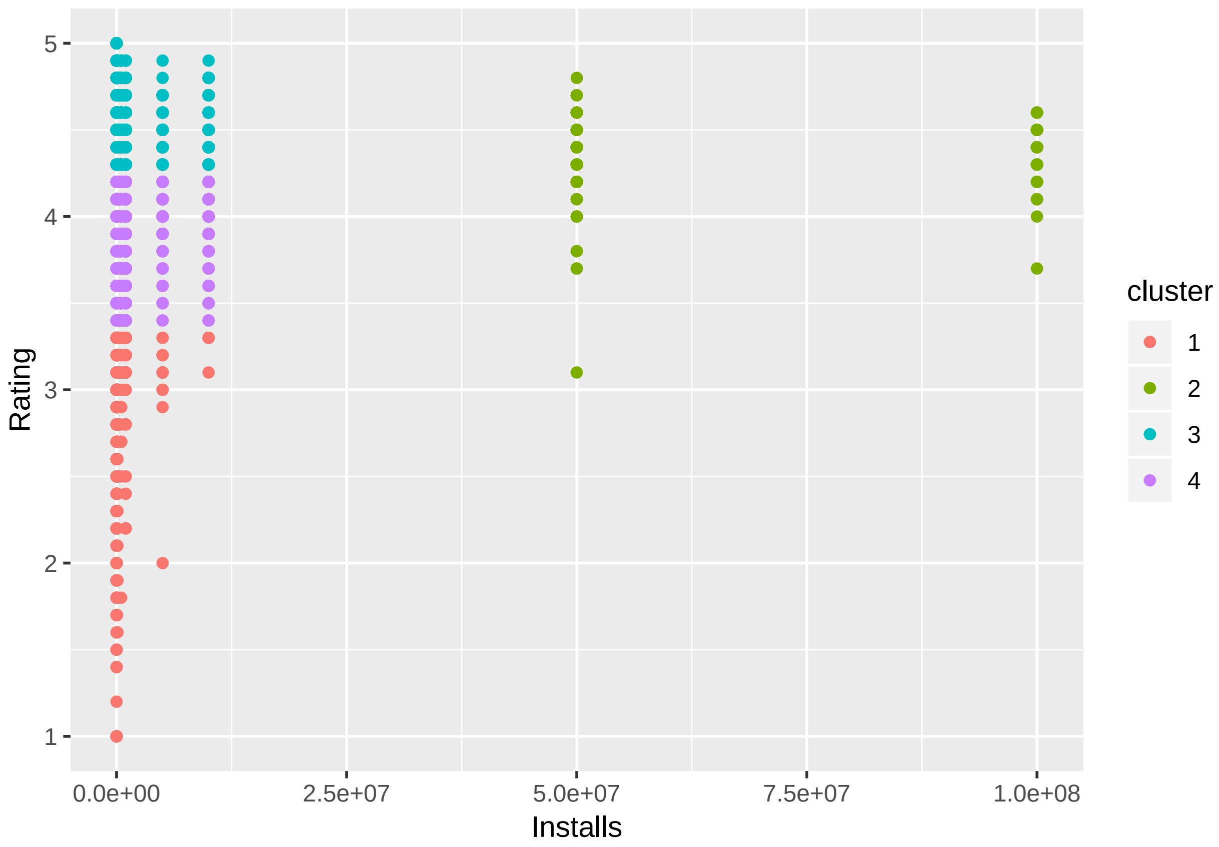
These 3 cluster models attempt to cluster ratings with various other categories in our dataset. We found little predictive capability from these models. The only potentially useful clustered result in the Ratings vs. Installs, where a clear correlation is visible in the cluster plot.

```r
library(ggplot2)
library(cluster)

#Initialize random variable
set.seed(30)

d=na.omit(d) #omit NA values
d$Reviews <- as.numeric(d$Reviews)
d$Rating <- as.numeric(d$Rating)
```

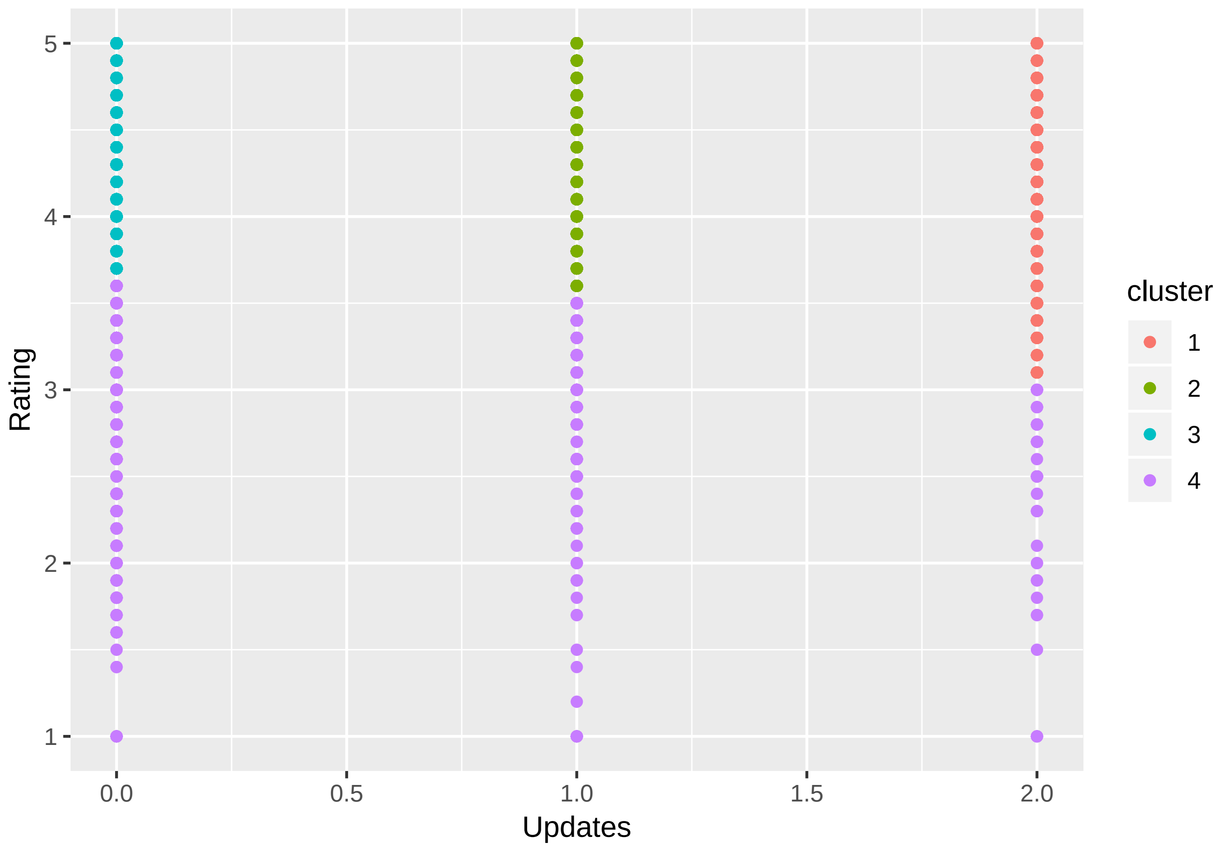
```
clusters<-kmeans(scale(d[,4:3]), 5, nstart=25)

d$cluster=as.factor(clusters$cluster)

ggplot(d, aes(x=Reviews, y=Rating, color=cluster)) +geom_point()
```



```
d1<- subset(d, select=c("Rating", "Installs"))
clusters2<- kmeans(scale(d1), 4, nstart=25)
d1$cluster=as.factor(clusters2$cluster)
ggplot(d1, aes(x=Installs, y=Rating, color=cluster)) +geom_point()
```



```
d2<- subset(d, select=c("Rating", "Updates"))
View(d2)
clusters3<- kmeans(scale(d2), 4, nstart=25)
d2$cluster=as.factor(clusters3$cluster)
ggplot(d2, aes(x=Updates, y=Rating, color=cluster)) +geom_point()
```


Rating
5
4
3
2
1
0.0
0.5
1.0
1.5
2.0
Updates
cluster
1
2
3
4