iOS App Rating Dataset

Group 2:
Diogo Moura, Rustam Zayanov, Mustafa Khalil, Silvestre Pires

Overview

1. Prediction of app popularity + App feature selection: A concern of
    - Business owners
    - App developers

2. Dataset from Kaggle
    - https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps
    - 7,197 records
    - 17 variables.

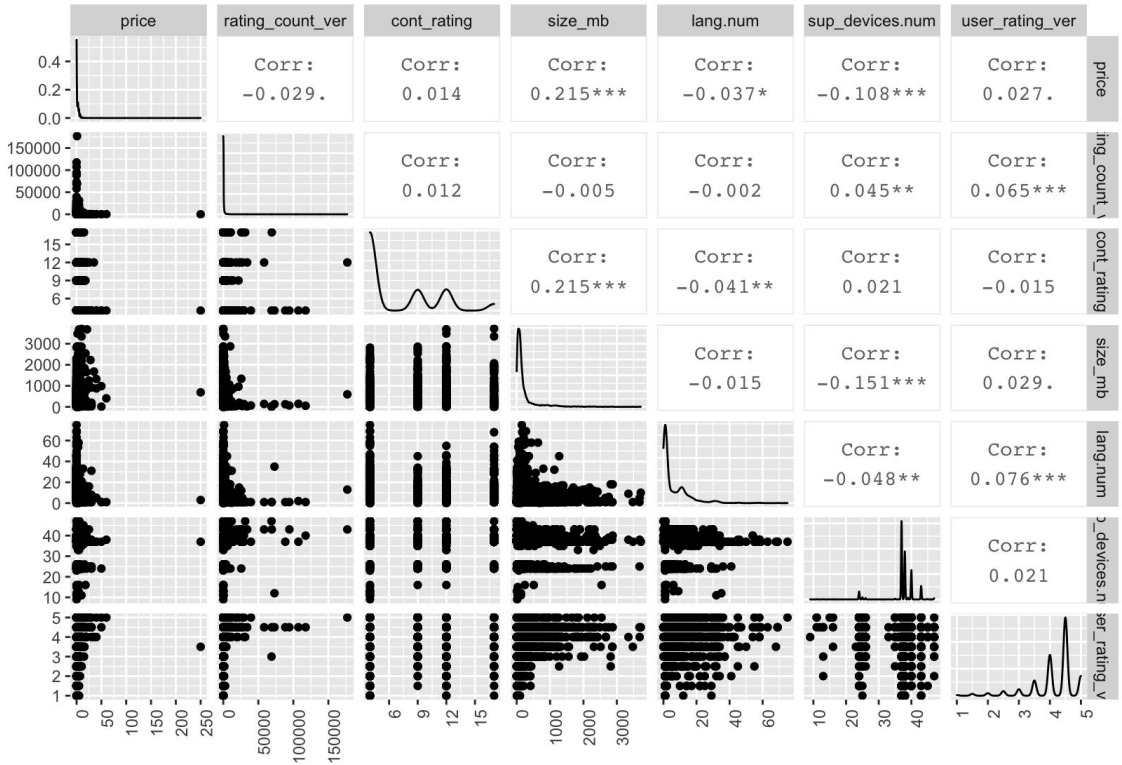
Exploratory Data Analysis

Some Variables:

1. Track_name ---- The application name.
2. Size_bytes ---- Application size in bytes.
3. Currency ---- Currency of the application price. All values are equal to "USD".
4. Price ---- Application price in the Store. E.g. 2.99.
5. Rating_count_tot ---- Total number of user ratings, for all app versions.
6. Rating_count_ver ---- Number of ratings for the current app version.
7. User_rating ---- Average user rating for all versions, from 0 to 5.
8. User_rating_ver ---- Average user rating for the current version, from 0 to 5.
9. Ver ---- The current version number of an app. E.g. "4.0.4"

- The class variable is User_rating
  - but it was changed to User_rating_is_good
  - Binary Value: 0 if less than 4, 1 otherwise

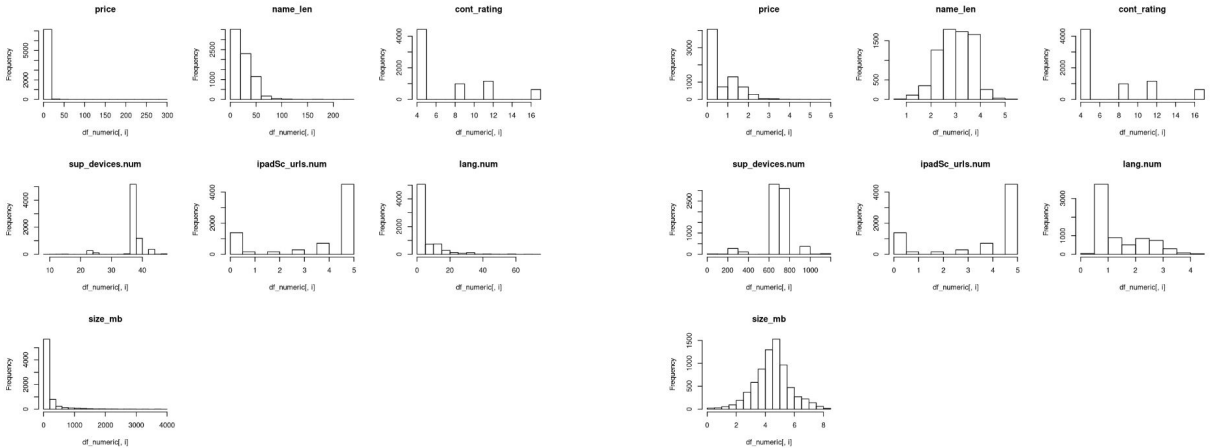
Exploratory Data Analysis



- Variables not normal
- Not much of correlation

Variables Transformation

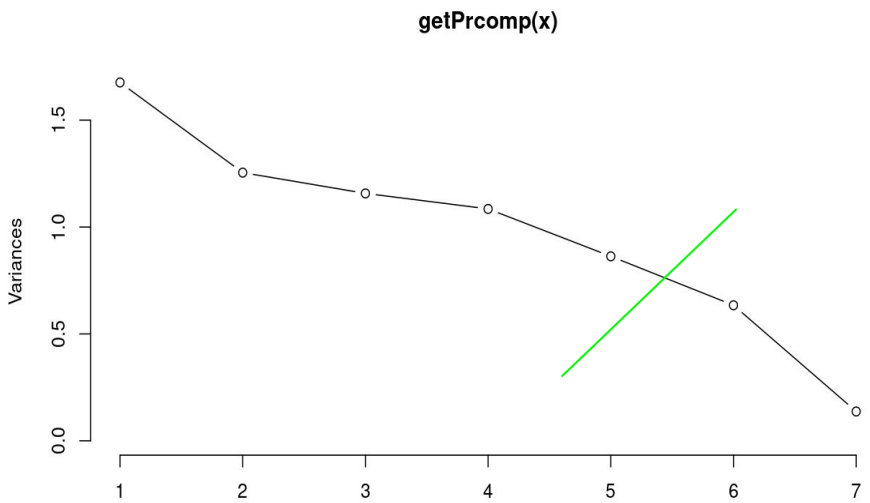
1. Marking apps with no ratings as "bad"
2. Numerical Variables: Box-Cox Transformation
3. Squaring *sup_devices_num*
4. Log transformation on *price*, *name_len*, *lang.num*, and *size_mb*.



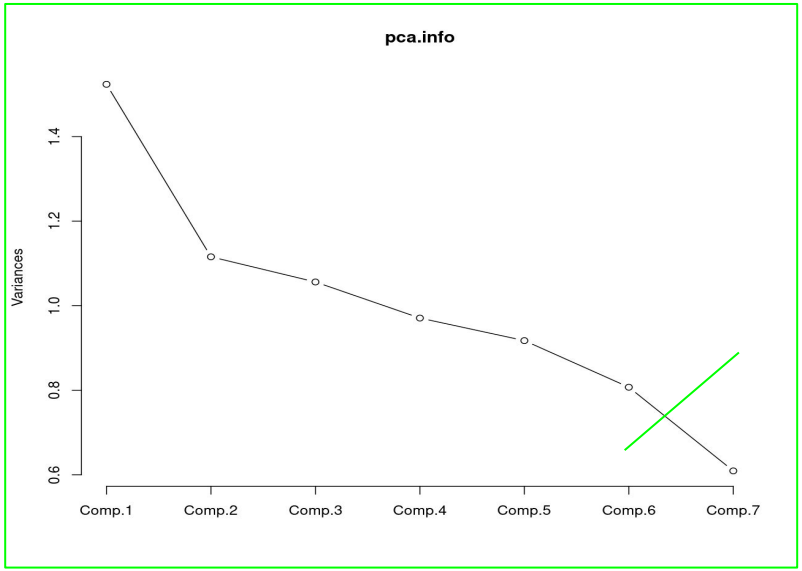
Before and after Box-Cox transformation

Dimensionality Reduction

5 Components for robust PCA



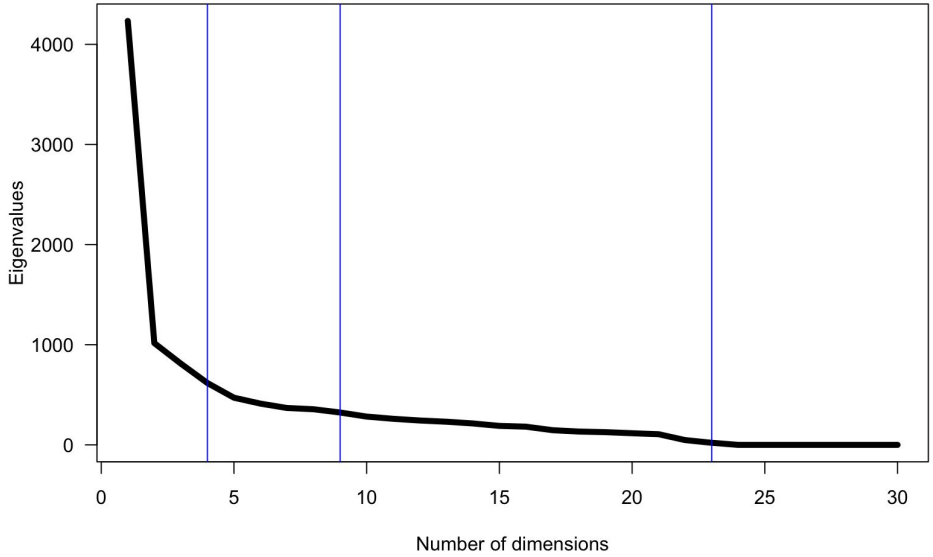
6 Components for classic PCA



Dimensionality Reduction

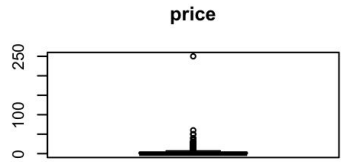
Categorical Variables

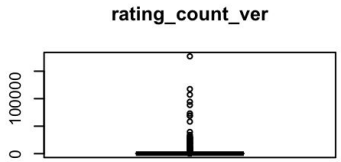
- Multidimensional Scaling (MDS)
- 80% of variance is described by 9 PCs
- The number of above-average eigenvalues is 2

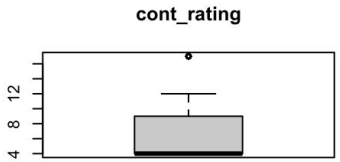


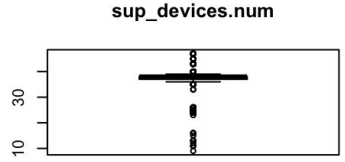
Outliers Analysis & Removal

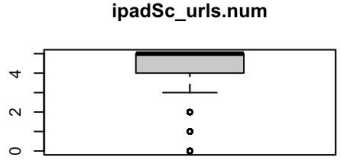
Univariate Detection

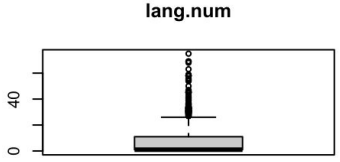


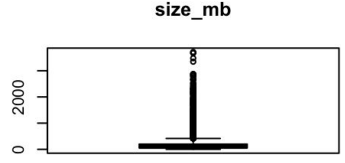












Outliers Analysis & Removal

Score-based Detection and Orthogonal Distances

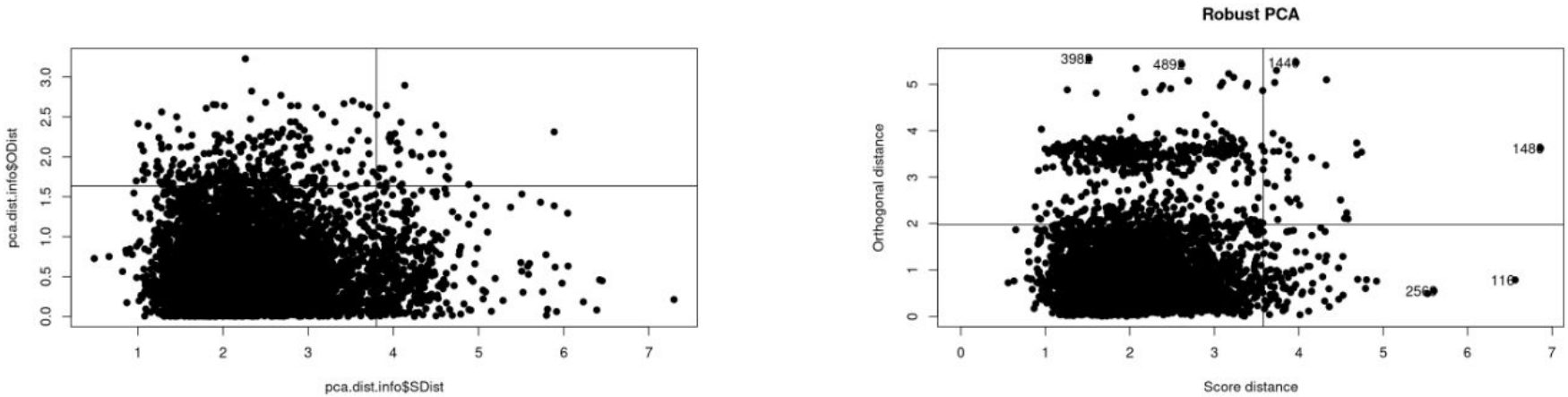


Figure 7: Outlier plots for classic and robust PCA

Clustering: Results

The Chosen →

#	Dataset	Outliers	Method	Bal.Avg	k	Silh	CH
1	Numeric+Cat	No	Single	.5	2	.33	3
2	Numeric+Cat	No	Complete	.5	2	.32	58
3	Numeric+Cat	No	PAM	.6	2	.34	2524
4	Numeric	No	Complete	.5	3	.7	562
5	Numeric	No	PAM	.6	8	.17	977
6	Rob.PCA+Cat	No	Complete	.5	2	.32	58
7	Rob.PCA+Cat	No	Average	.5	2	.32	58
8	Rob.PCA+Cat	No	Ward	.5	2	.06	393
9	Rob.PCA+Cat	No	PAM	.59	4	.03	353
10	Rob.PCA+MDS	No	Complete	.59	10	.1	488
11	Rob.PCA+MDS	No	PAM	.63	4	.18	1132
12	Rob.PCA	No	Complete	.6	4	.13	969
13	Rob.PCA	No	PAM	.61	10	.24	1354
14	Rob.PCA+Cat	Yes	Complete	.59	10	.41	794
15	Rob.PCA+Cat	Yes	PAM	.58	4	.33	1344
16	Rob.PCA+MDS	Yes	Complete	.5	2	.19	140
17	Rob.PCA+MDS	Yes	PAM	.62	6	.17	1130
18	Rob.PCA	Yes	Complete	.5	2	.32	94
19	Rob.PCA	Yes	PAM	.6	9	.23	1497

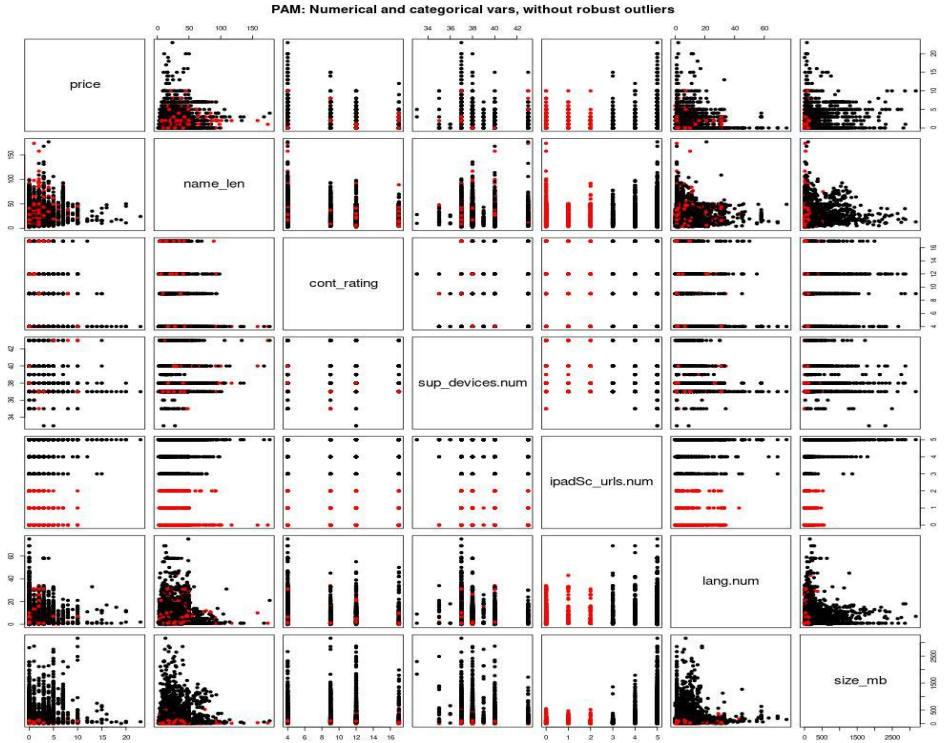
Clustering: Interpretation

Results of the best clustering

Cluster	App is good	Not good
1	3371	1570
2	611	916

No Clear separation in bad/good apps in each cluster

Clustering: Interpretation



Clear separation by Screenshot count

Classification

Without dimensionality reduction

	Accuracy	Balanced Acc	Sensitivity	Specificity	Avg F1-score	PPV	NPV
LDA	0.701	0.664	0.837	0.491	0.668	0.718	0.660
RDA	0.597	0.501	0.945	0.057	0.420	0.608	0.400
SVM	0.699	0.659	0.842	0.479	0.663	0.714	0.661
Neural Network	0.694	0.659	0.819	0.499	0.663	0.717	0.640
Random Forest	0.689	0.661	0.789	0.534	0.664	0.724	0.620
KNN	0.675	0.638	0.809	0.468	0.641	0.702	0.612

Classification

With robust PCA and MDS variable transformation

	Accuracy	Balanced Acc	Sensitivity	Specificity	Avg F1-score	PPV	NPV
LDA	0.693	0.652	0.842	0.462	0.655	0.708	0.654
RDA	0.701	0.659	0.852	0.465	0.662	0.712	0.670
SVM	0.701	0.660	0.852	0.469	0.664	0.713	0.671
Neural Network	0.693	0.651	0.845	0.457	0.654	0.707	0.656
Random Forest	0.683	0.654	0.788	0.521	0.657	0.718	0.613
KNN	0.693	0.662	0.807	0.516	0.665	0.721	0.633

Classification

Ensemble decisions metrics with the dataset with no PCA and with One-hot Encoding

	Accuracy	Balanced Acc	Sensitivity	Specificity	Avg F1-score	PPV	NPV
Average	0.713	0.672	0.865	0.478	0.676	0.720	0.696
Weighted Average	0.711	0.673	0.852	0.494	0.677	0.723	0.682
Majority	0.707	0.665	0.861	0.468	0.669	0.715	0.685

Results are better than non-ensemble methods.

Classification

Classification into Clusters

	Accuracy	Balanced Accuracy	F1-score
LDA	0.971	0.965	0.961
RDA	0.971	0.965	0.961
SVM	0.981	0.972	0.974
Neural Networks	0.989	0.989	0.985
Random Forest	0.971	0.959	0.961
KNN	0.982	0.975	0.975

Conclusion

1. Overview:
    - Variables not normal -- transformation was necessary
    - Low variable correlation, dimensionality can hardly be reduced

2. Classification:
    - Classifier: Ensemble method
    - Balanced accuracy: 67%
    - Data: transformed numeric variables + one-hot encoded categorical variables

3. Clustering:
    - Method: PAM, k = 2
    - Balanced accuracy 60%, good Silhouette and CH scores
    - Near-perfect classification with NNs
    - Not recommended for classification

4. Future Work: Deal with The problem of Imbalance between Good and Bad apps.