GOOGLE PLAY APPS

Campeny, Eloi
Chriki, Fatima Zohra
Dai, Zhongkai
González, Victor
Moure, Ximena
Xu, Ange

TABLE OF CONTENTS

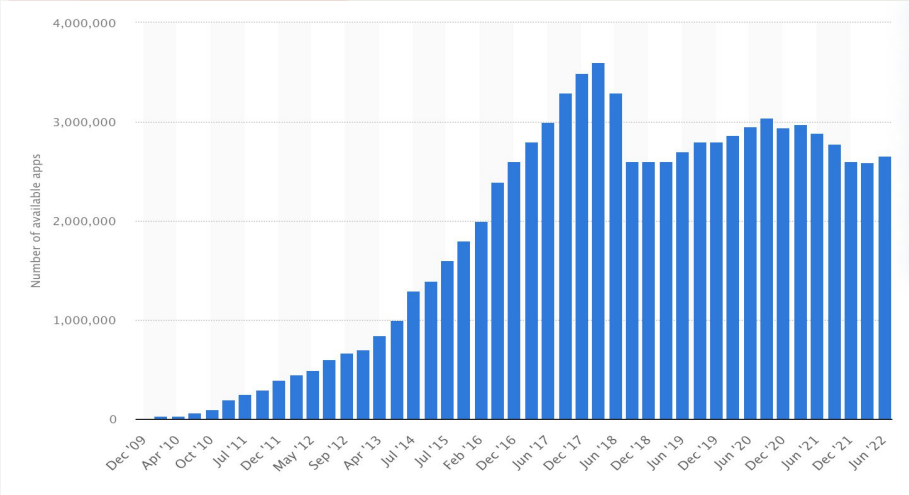
01

Motivation & Problem definition

Motivation & Problem definition

Google Play

+2500 daily additions



Analyze which factors can influence the number of downloads and the rating of an app

02

Data source presentation

Data source presentation

- Data taken from Kaggle
- Collected in 2021
- Original dataset: 2.312.944 observations
- Reduced dataset: 20.000 observations

03

Data structure and metadata description

Data structure and metadata description

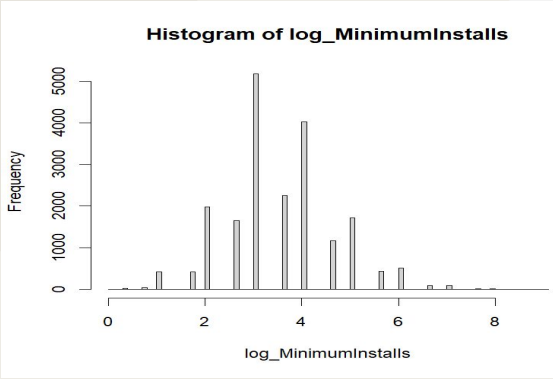
- **24 variables**
  - **5 numerical:** Rating, Rating.Count, Minimum.Installs, Maximum.Installs and Price
  - **4 binary:** Free, Ad.Supported, In.App.Purchases and Editors.Choice
  - **15 categorical:** Category, Installs, Size, Released, Last.Updated, Content.Rating,Scrapped.Time,Developer.Email, Developer.Website, etc.
- 10911 **missing values** (2.27%)
- Variables eliminated and transformed during preprocessing

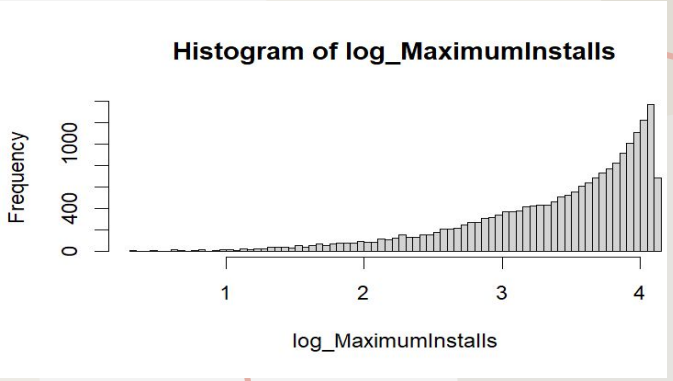
04

Preprocessing

Preprocessing

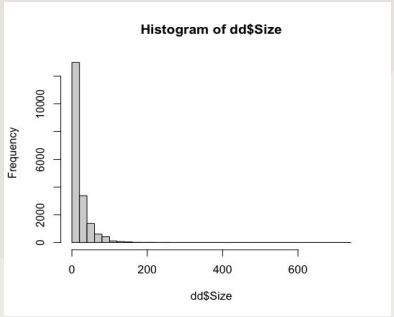
- **Feature selection**
  - **Unique value for each observation**
    - Developer.Website, Developer.Email, Developer.Id, App.Id, Privacy.Policy
  - **Unique value for all observations**
    - Editors.Choice
  - **Highly correlated**
    - Installs , Minimum.Installs, Maximum.Installs → Installs

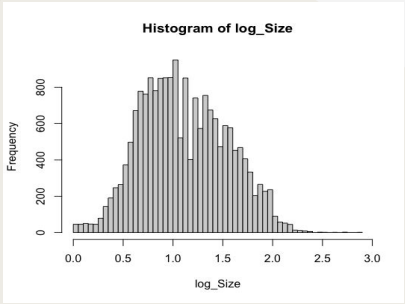


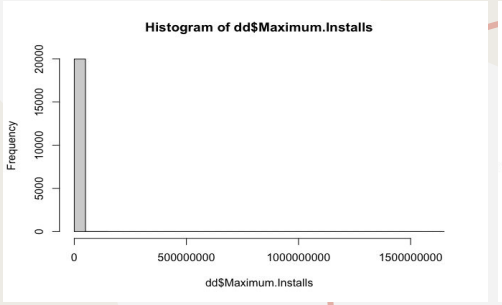


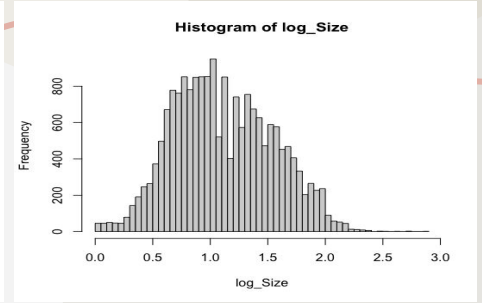
Preprocessing

- **Derivation of new variables**
  - `Scraped.Time, Last.Updated` → `DaysLastUpdate`
  - `Scraped.Time, Release` → `ReleaseDays`
  - `App.Name` → `AppNameLen`
- **Transformations**
  - Categorial `Size` (GB, MB, KB) → Numerical
  - `Size`, `Installs` and `Rating.Count` → Logarithmic transformation
  - Reduced modalities for `Category` and `Minimum.Android`









Preprocessing

- **Segmentation of population**
  - **Paid vs Free Apps**
  - **Kolmogorov-Smirnov test (KS test)**
    - Divided dataset: Paid apps and Free Apps
    - Performed the test for each numerical variable
  - **Result:** p-value < 0.05 for almost every variable
  - **Conclusion:** There is more than one population

Missing data treatment

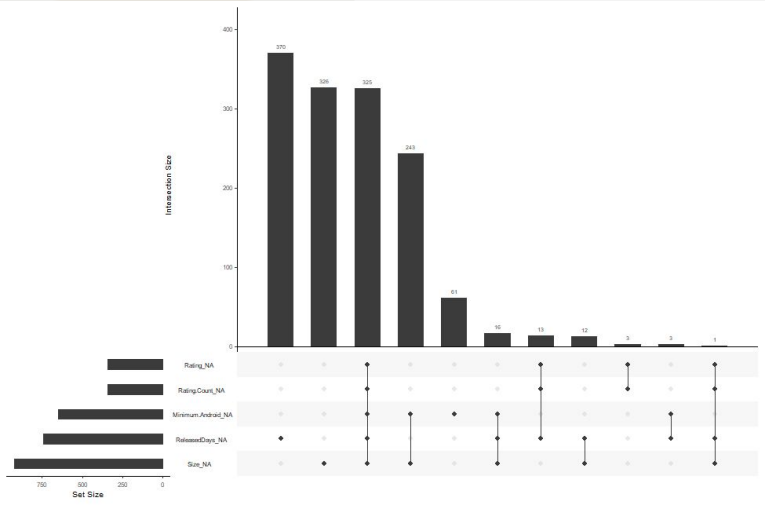
- **Identifying missing values**
  - Identify variables affected
  - Identify type of missing data
    - Little Test result: p=0
    - Analysis of missing data → MAR

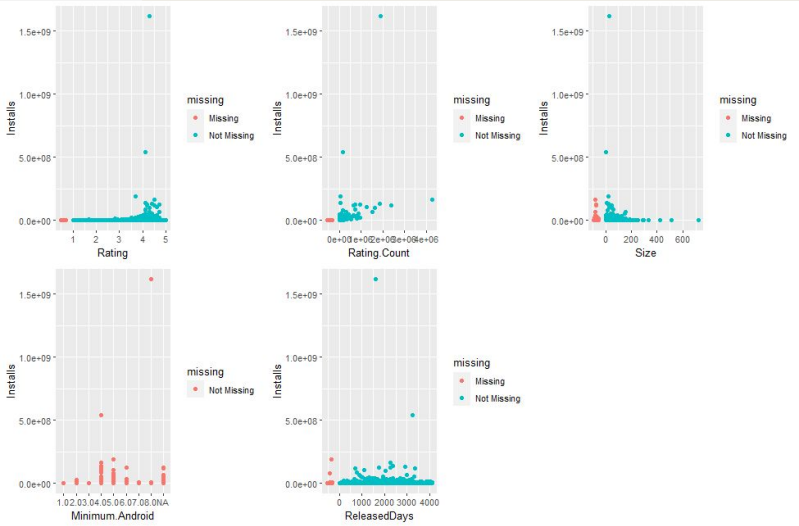
$Continuous

	label	var_type	n	missing_n	missing_percent	mean
Rating	Rating	<dbl>	19227	342	1.7	4.1
Rating.Count	Rating.Count	<int>	19227	342	1.7	3342.7
Size	Size	<dbl>	18646	923	4.7	20.3
DaysLastUpdate	DaysLastUpdate	<dbl>	19569	0	0.0	562.5
ReleasedDays	ReleasedDays	<dbl>	18829	740	3.8	1195.8
AppNameLen	AppNameLen	<int>	19569	0	0.0	23.1
Installs	Installs	<int>	19569	0	0.0	431313.5

$Categorical

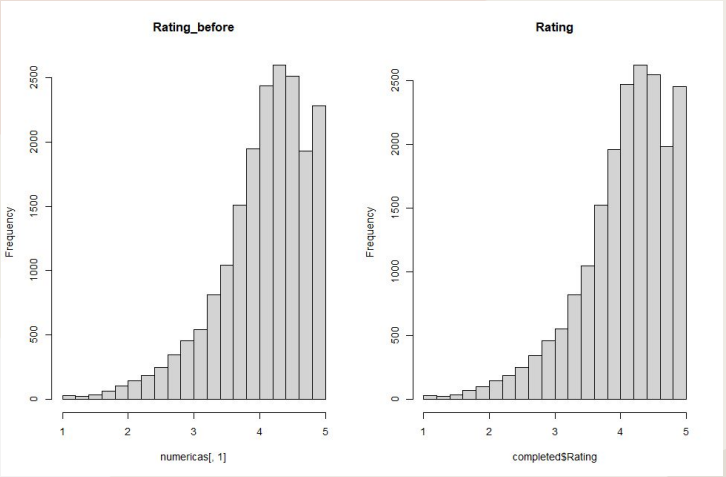
	label	var_type	n	missing_n	missing_percent	levels_n
Category	Category	<fct>	19569	0	0.0	6
Minimum.Android	Minimum.Android	<fct>	18921	648	3.3	8
Content.Rating	Content.Rating	<fct>	19569	0	0.0	6
Ad.Supported	Ad.Supported	<fct>	19569	0	0.0	2
In.App.Purchases	In.App.Purchases	<fct>	19569	0	0.0	2

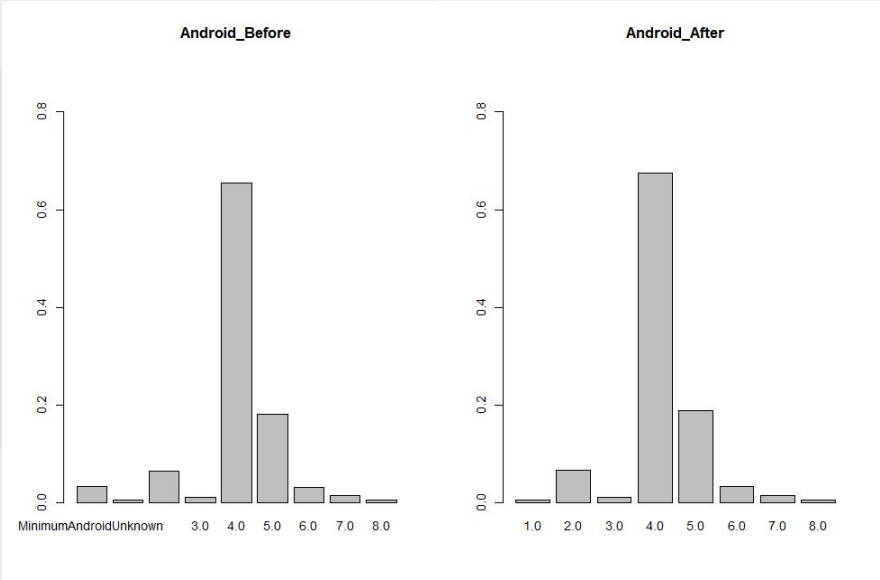


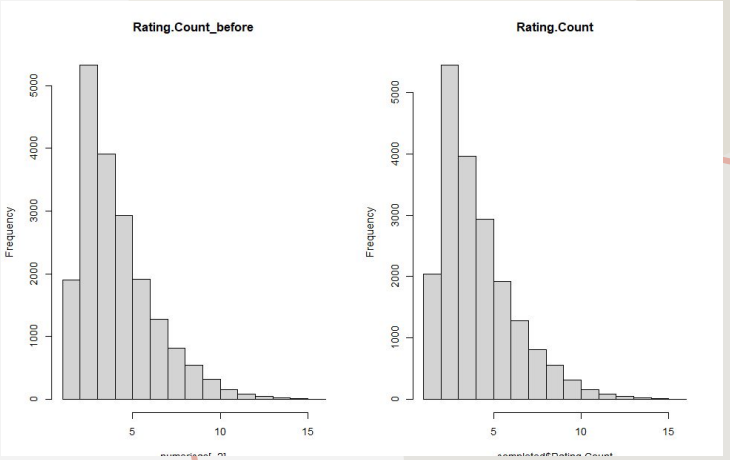


Missing data treatment

- **Treating missing values**
  - MICE with numerical data
  - MICE with categorical data

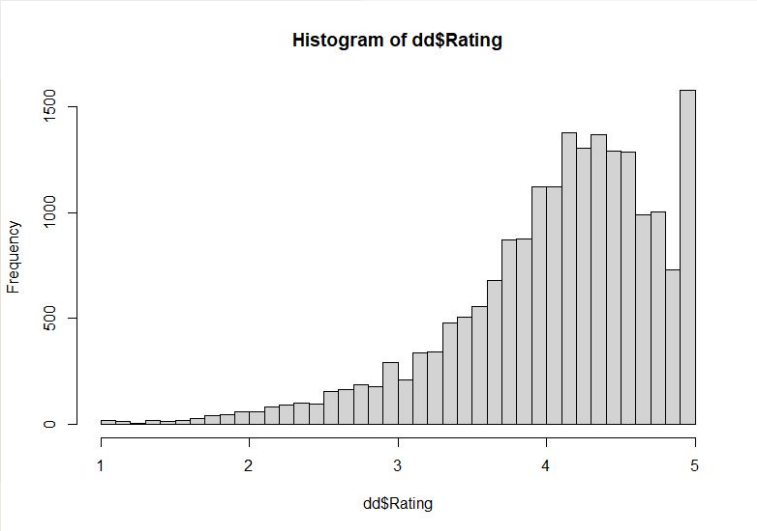


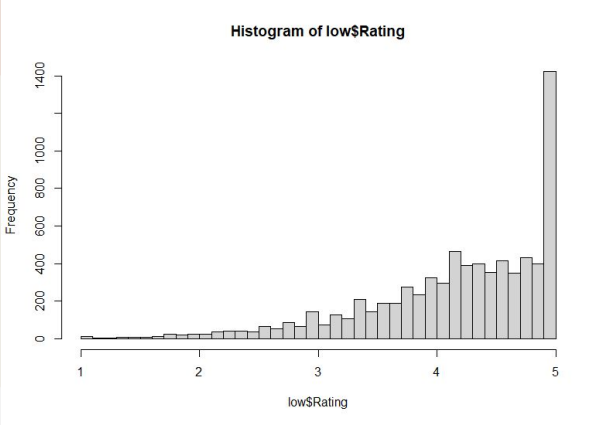


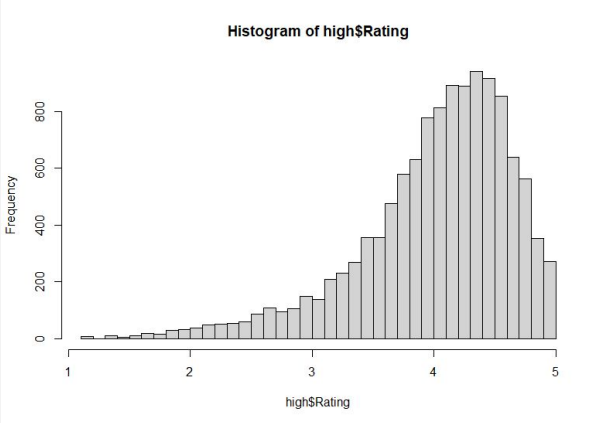


Missing data treatment

- **Univariate detection of outliers**
  - Rating analysis
    - Kolmogorov–Smirnov test and plot
  - Different populations, keep Rating.Count > 20

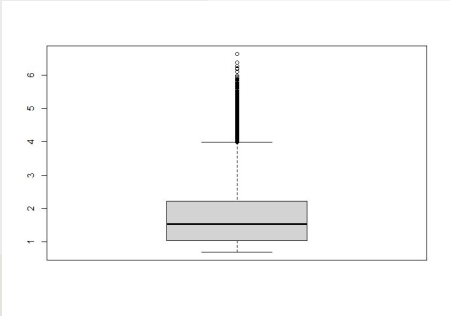


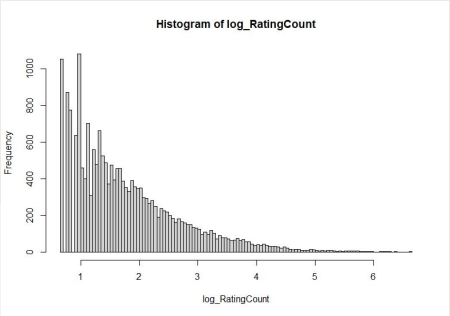




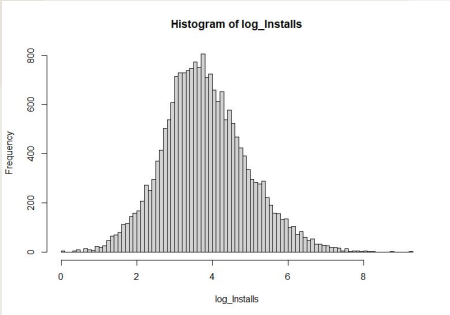
Missing data treatment

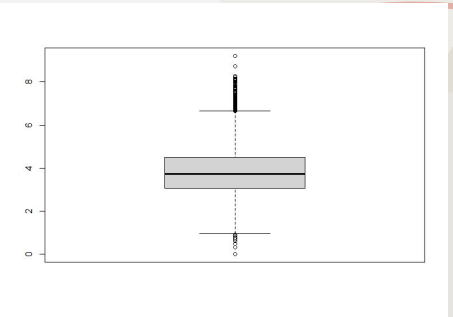
- **Univariate detection of outliers**
  - Rating.Count





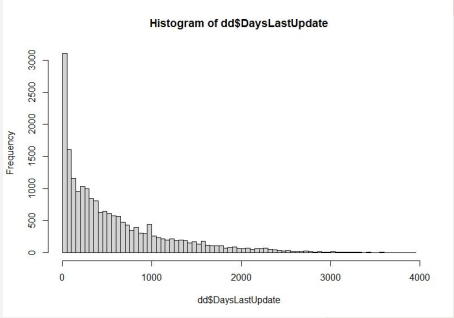
  - Installs

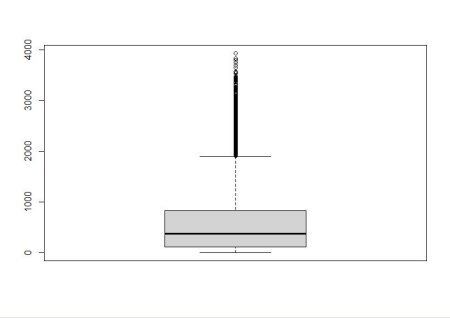




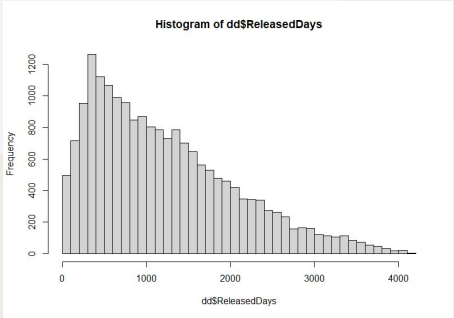
Missing data treatment

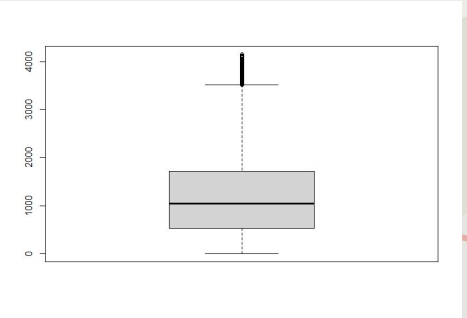
- **Univariate detection of outliers**
  - DaysLastUpdated





  - ReleasedDays





- **Multivariate Outliers**
  - Mahalanobis distance

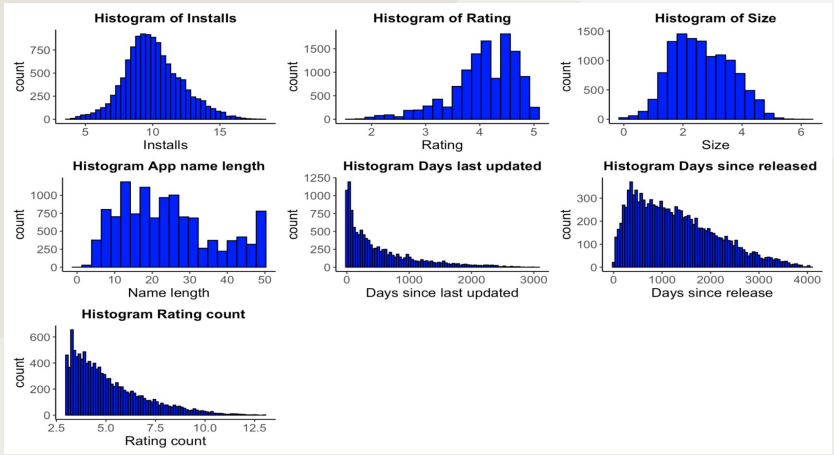
05

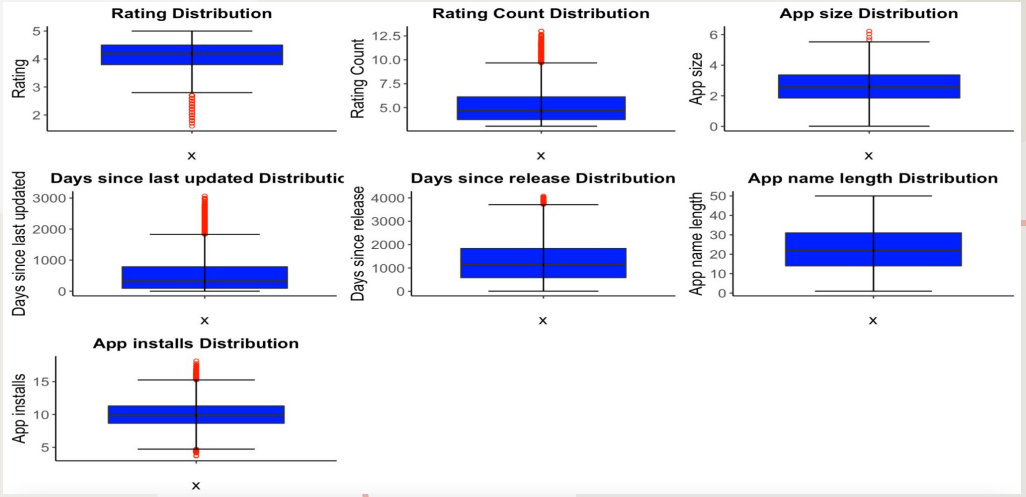
Basic initial descriptive statistics of preprocessed variables and conclusions

Univariate analysis

- **Univariate analysis**
  - Basic values, histograms and boxplots
    - Low Rating count respect Installs
    - Rating 4+
    - Name length 30-10

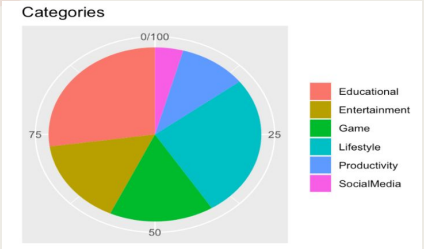
	Min	1st Q.	Median	Mean	3rd Q.	Max
Rating	1.600	3.800	4.200	4.078	4.500	5.000
Rating.Count	3.045	3.738	4.654	5.148	6.116	13.013
Size	0.01094	1.85630	2.56495	2.63017	3.36730	6.23637
DaysLastUpdate	0.0	91.0	325.0	534.4	787.0	3069.0
ReleasedDays	8	584	1146	1286	1837	4085
AppNameLen	1.00	14.00	22.00	24.18	31.00	50.00
Installs	3.664	8.661	9.863	10.048	11.303	18.165

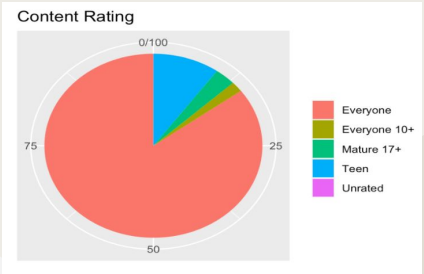


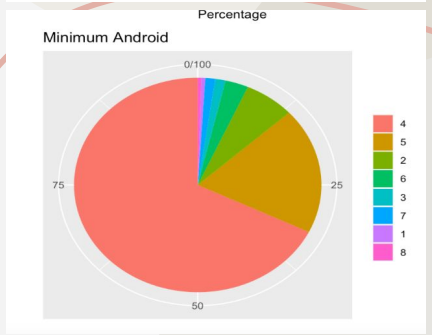


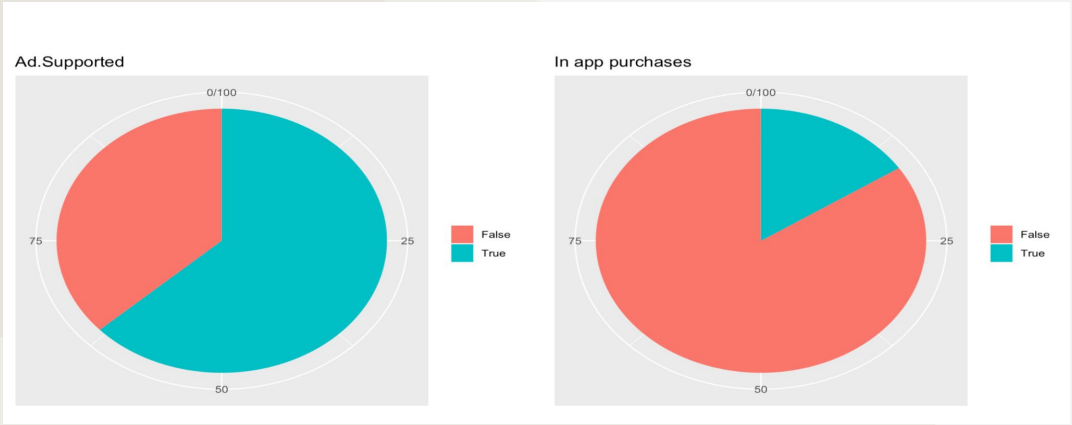
Univariate analysis

- **Univariate analysis**
  - Pie charts
    - In app purchases and adds
    - Educational › Lifestyle › Games › Entertainment
    - Everyone and Android 4+



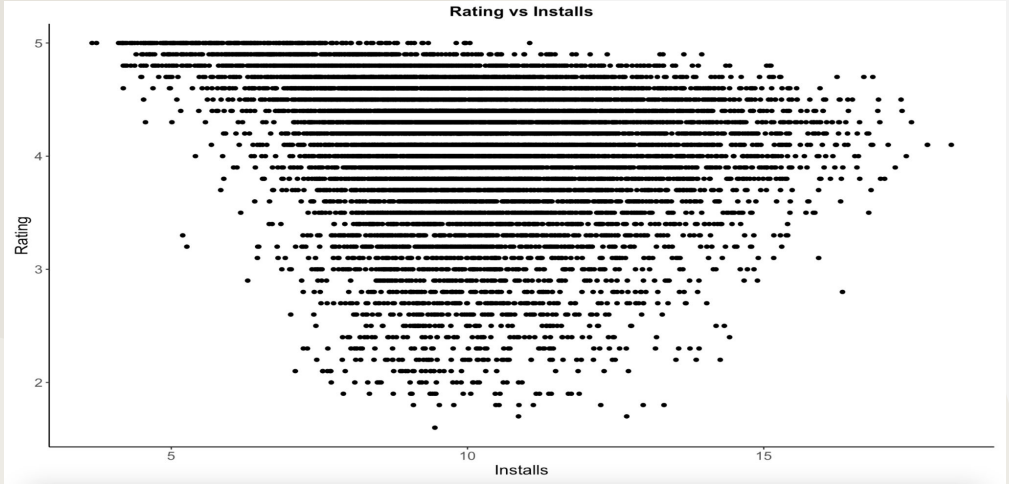


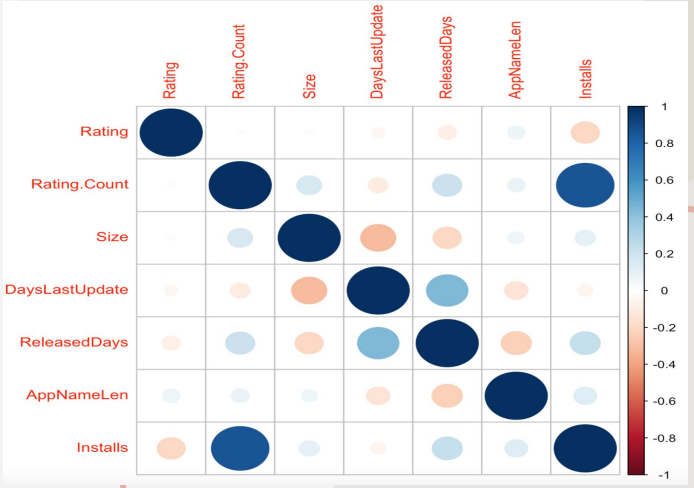




Bivariate analysis

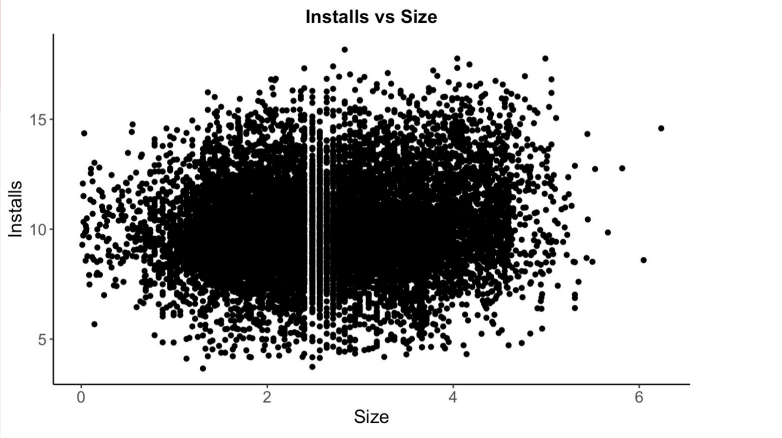
- **Bivariate analysis**
  - Correlation matrix
    - Positive correlation Installation+Rating.Count
    - Positive correlation DaysLastUpdate+ReleasedDays
  - Rating vs installs
    - Extremes higher scores



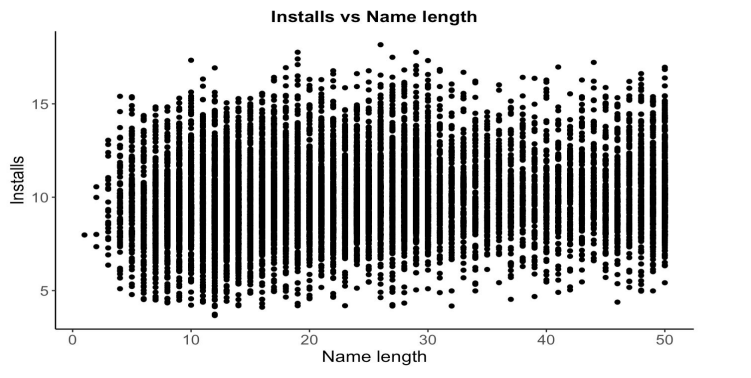


Bivariate analysis

- **Bivariate analysis**
  - Install vs size

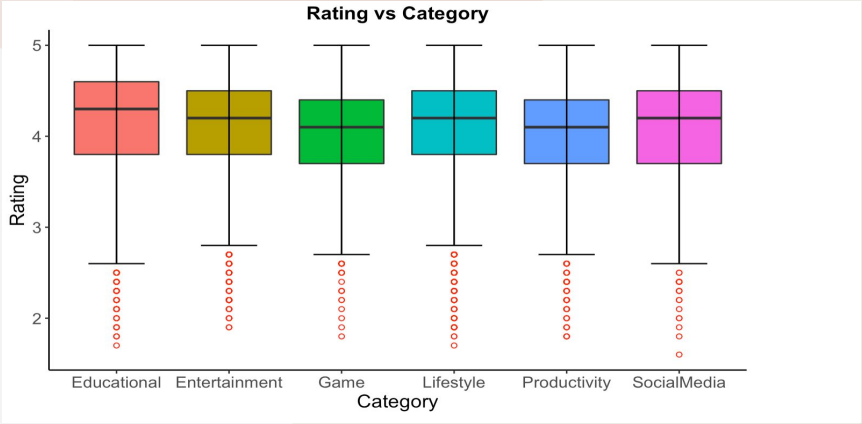


  - Installs vs Name length

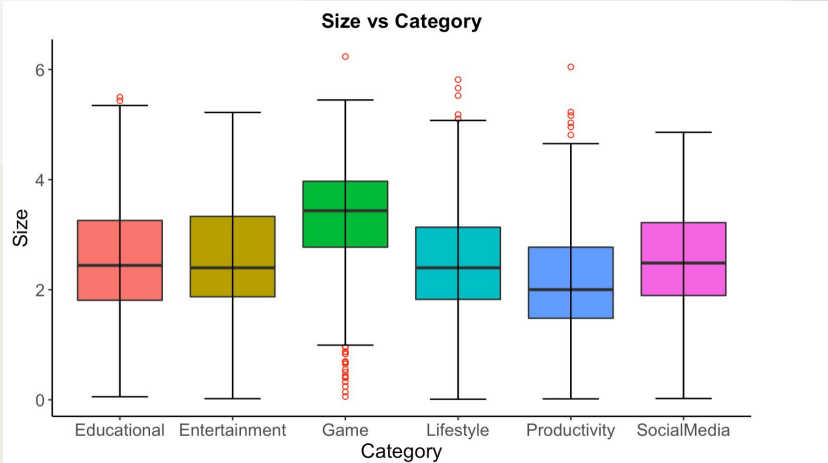


Bivariate analysis

- **Bivariate analysis**
  - Rating vs Category



  - Size vs Category

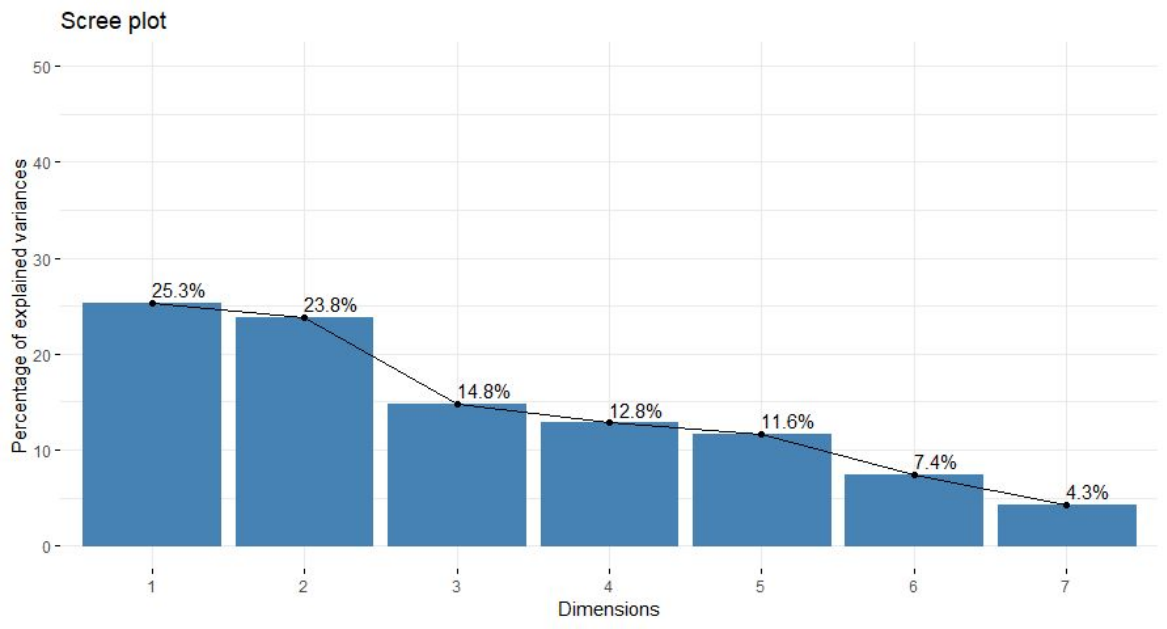


06

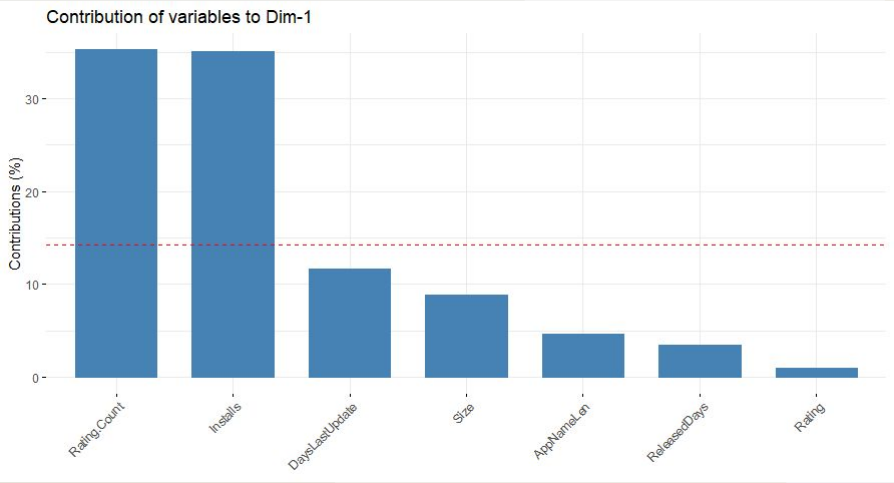
PCA for numerical variables

PCA: Variance retained by each component

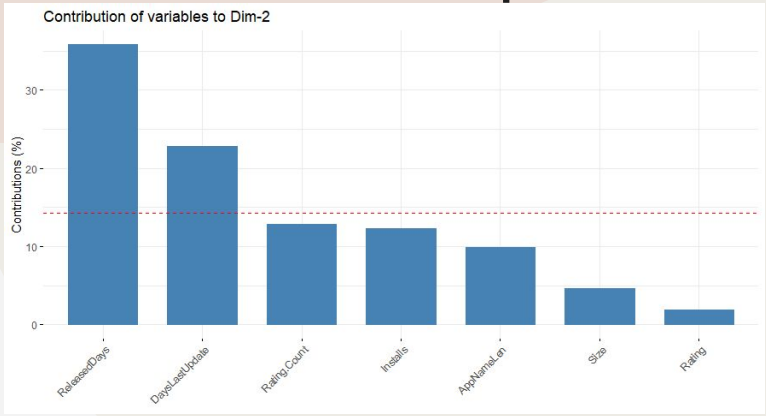
- 7 numerical variables
- The first three components satisfied with 63.91% of the total variance



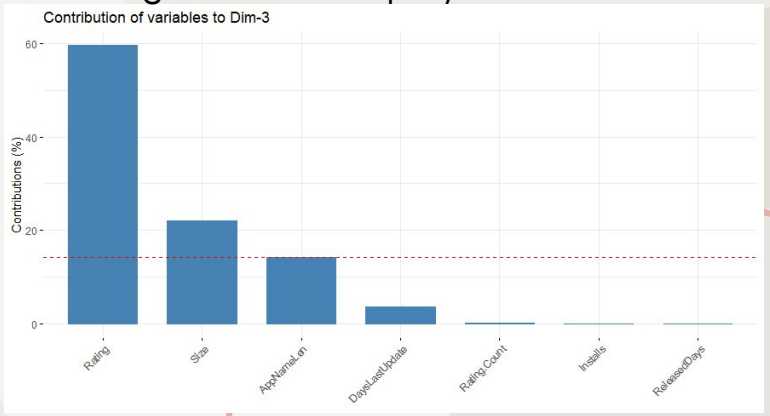
PCA: Contributions of variables to components



Meaning of PC1: Popularity

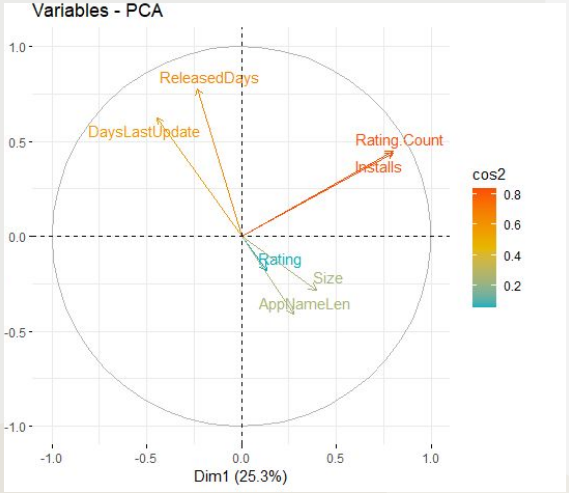


Meaning of PC2: Antiquity

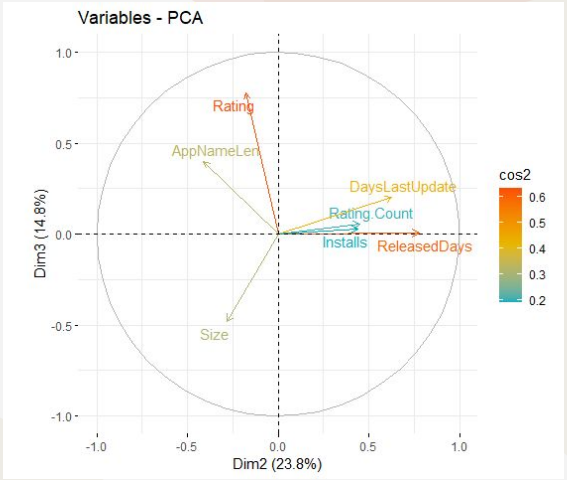


Meaning of PC3:Rating & app characteristics

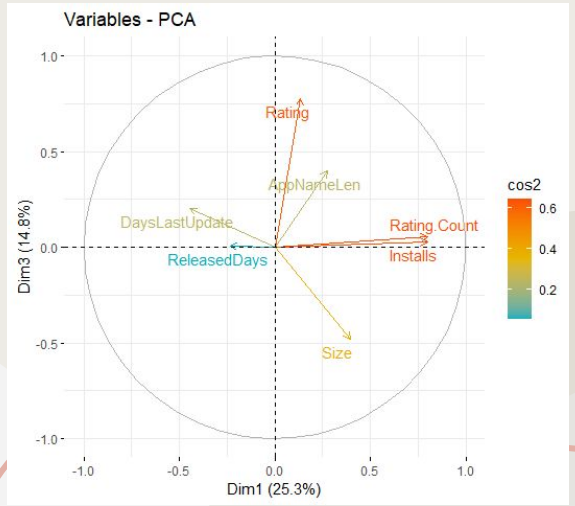
PCA: Correlation of variables to components



Dimensions 1-2



Dimensions 2-3



Dimensions 1-3

Rating.Count,***DaysLastUpdated*** and ***Installs*** are correlated with both PC1 and PC2

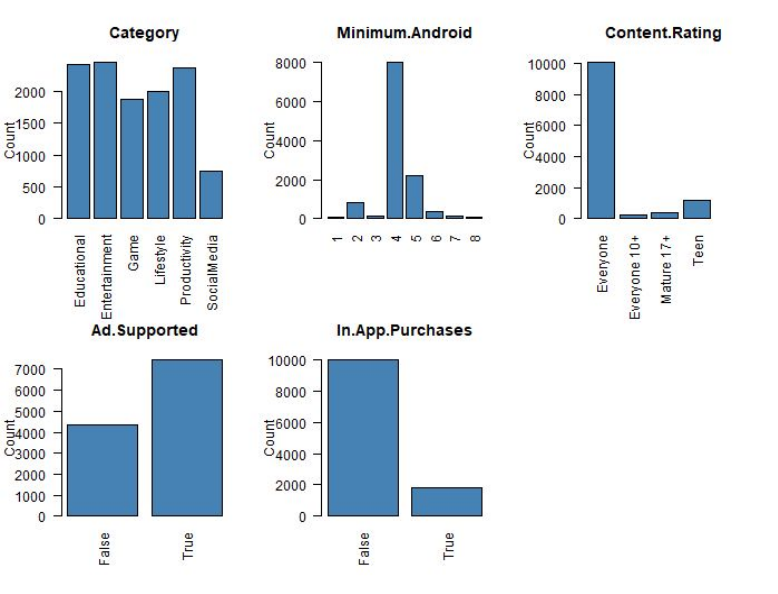
PCA: Biplots of individuals and variables

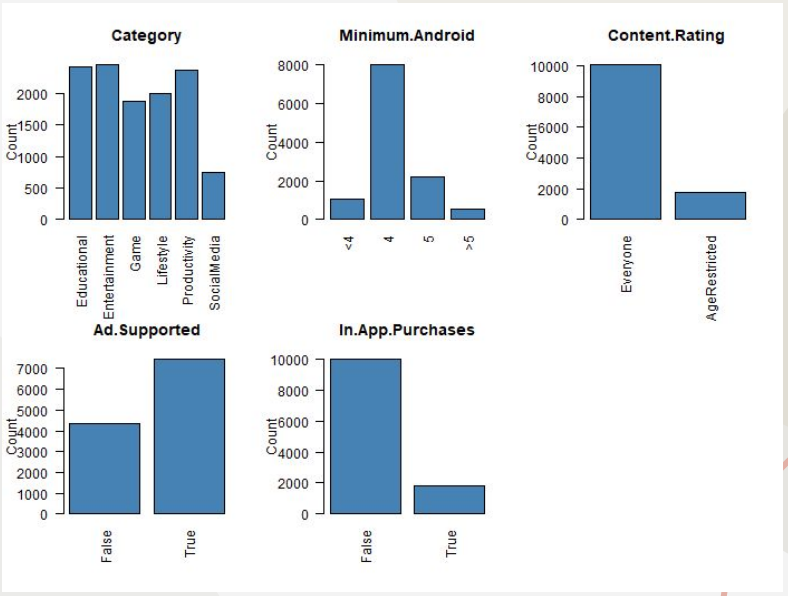
Dimensions 1-2

07

MCA of multiple qualitative variables

MCA: Detection of low frequency variable categories

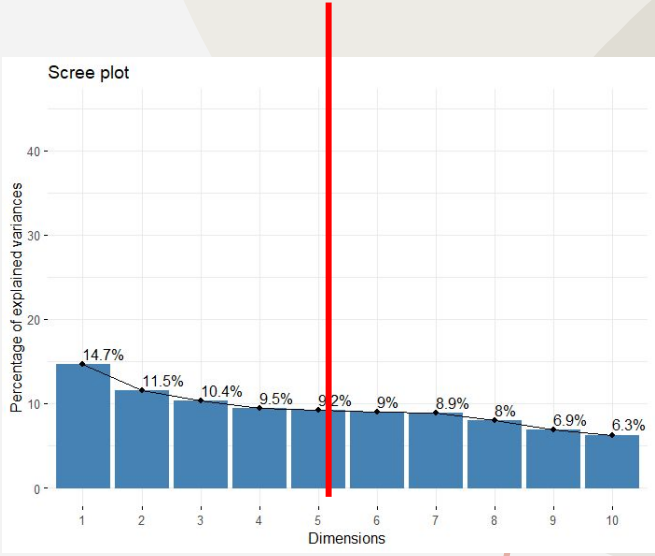




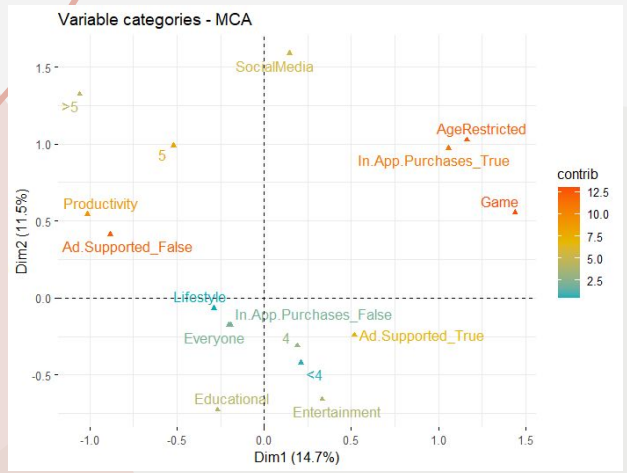
MCA: Eigen values

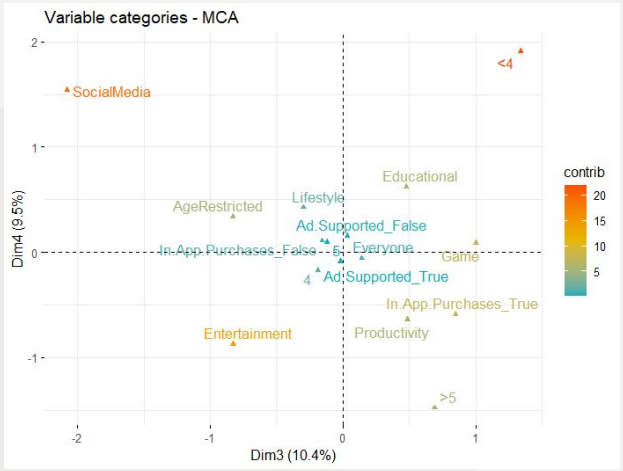
p = 5, 1/p = **0.2**

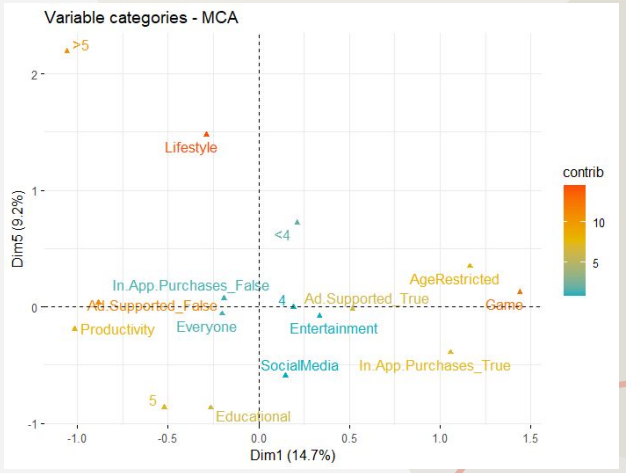
	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.3226002	14.663646	14.66365
Dim.2	0.2530929	11.504222	26.16787
Dim.3	0.2278697	10.357712	36.52558
Dim.4	0.2084731	9.476049	46.00163
Dim.5	0.2022985	9.195386	55.19701
Dim.6	0.1976014	8.981880	64.17889
Dim.7	0.1962563	8.920743	73.09964
Dim.8	0.1750459	7.956633	81.05627
Dim.9	0.1521349	6.915224	87.97150
Dim.10	0.1380674	6.275793	94.24729
Dim.11	0.1265597	5.752712	100.00000



MCA: Labeling the dimensions







Dim 1: Level of entertainment

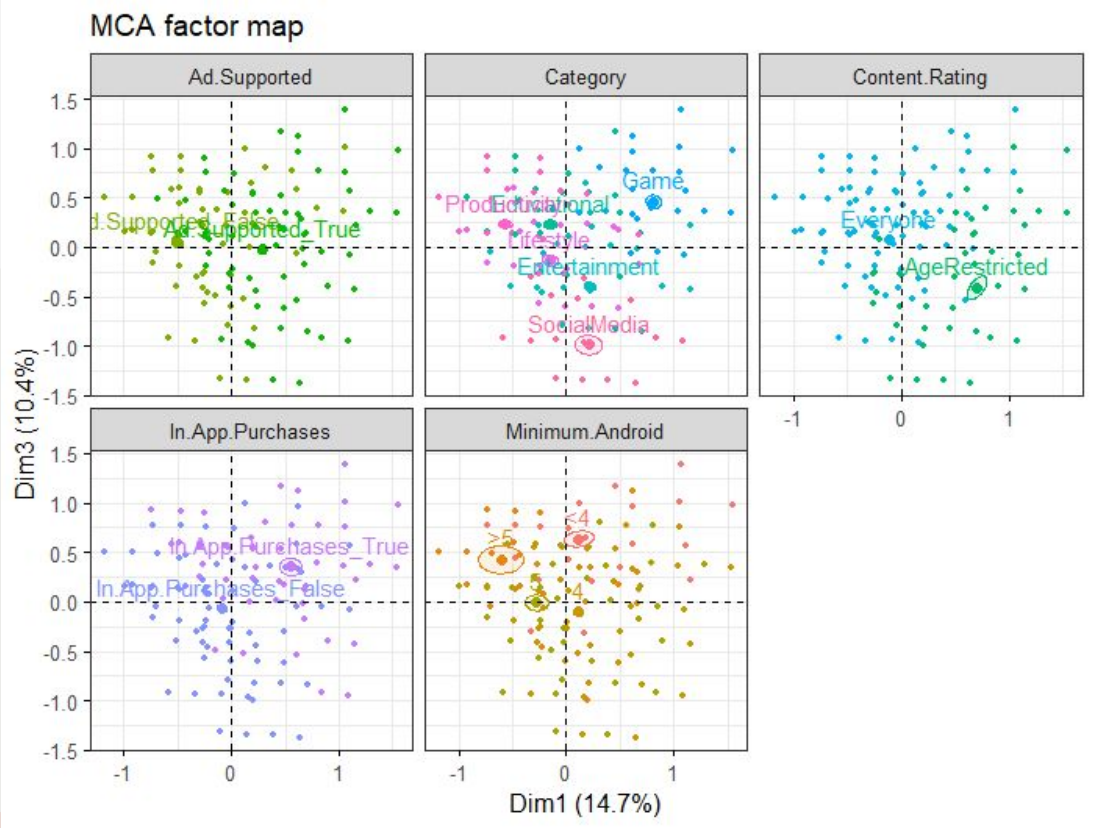
Dim 2: Level of procrastination

Dim 3: Level of companionship

Dim 4: Longevity

Dim 5: Helpfulness in a person's lifestyle

MCA: Individuals by groups

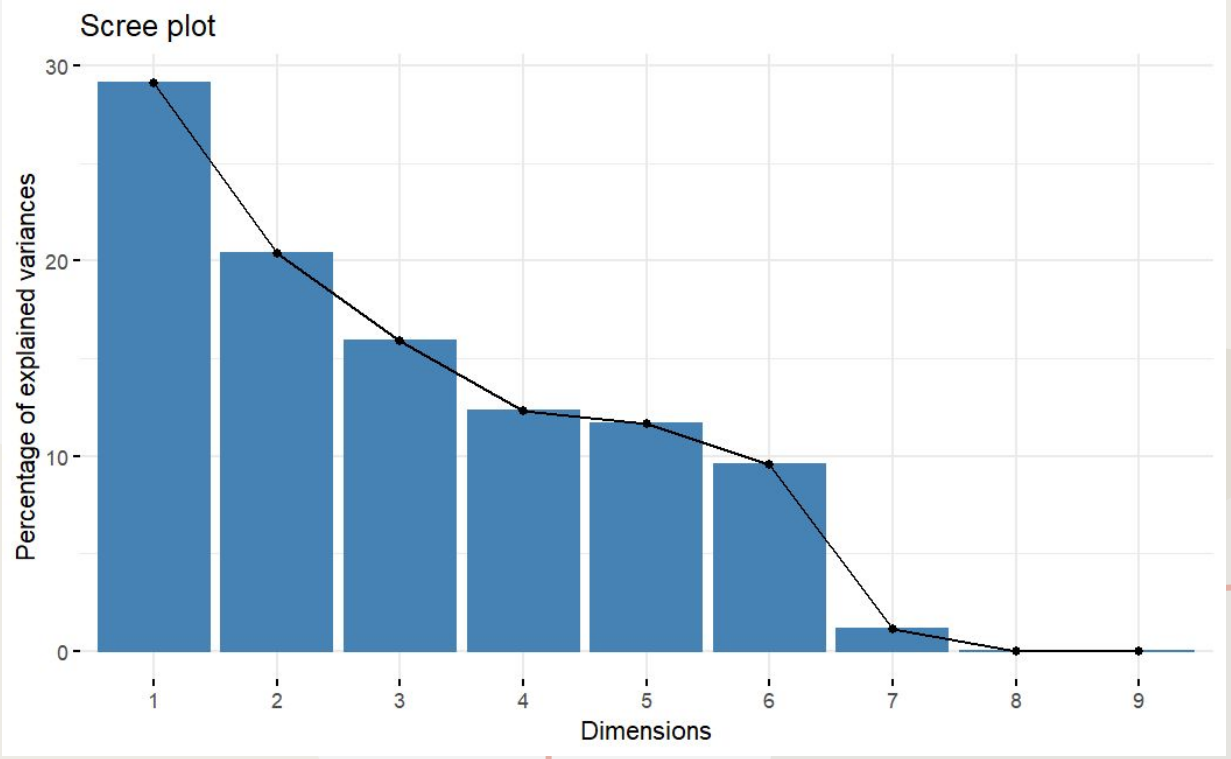


08

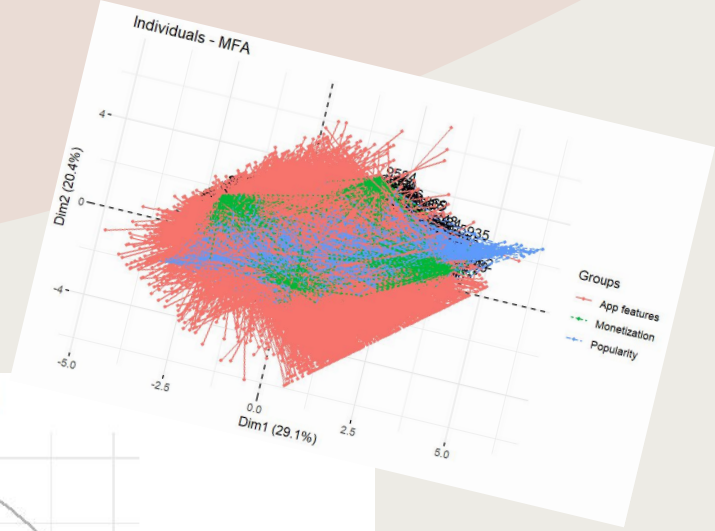
Multiple Factorial Analysis

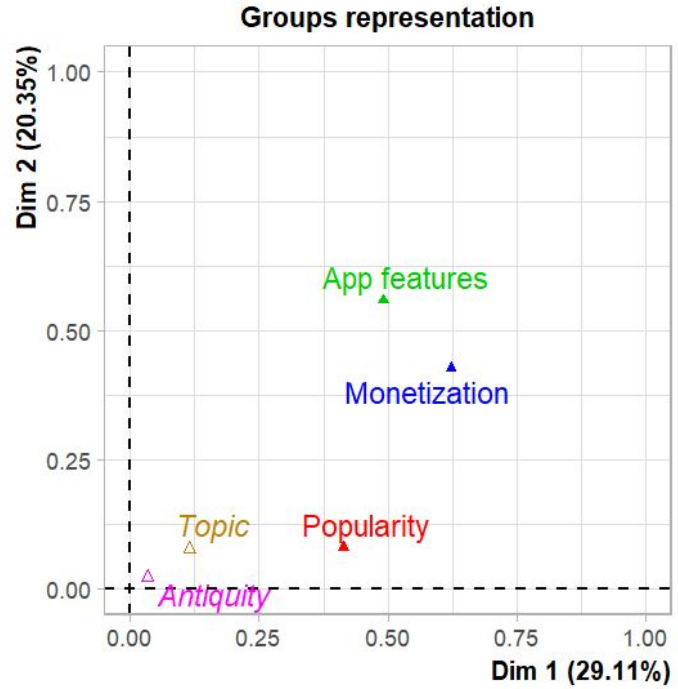
Groups & dimensions

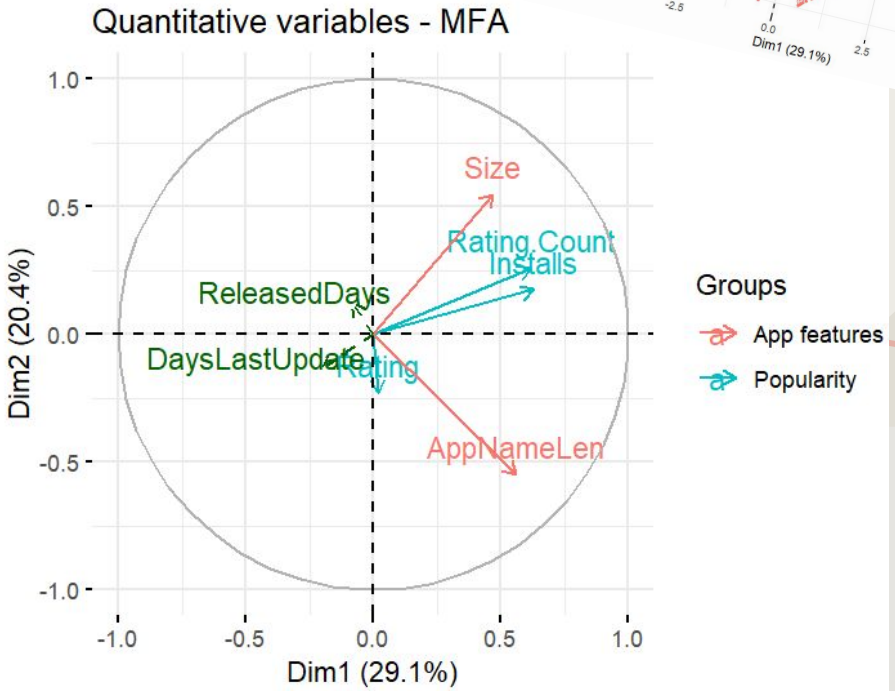
- **5 Groups**
  - **Antiquity**
  - **Popularity**
  - **App Features**
  - **Topic**
  - **Monetization**
- **3 DIMS**
- **65.35%** of cumulative variance



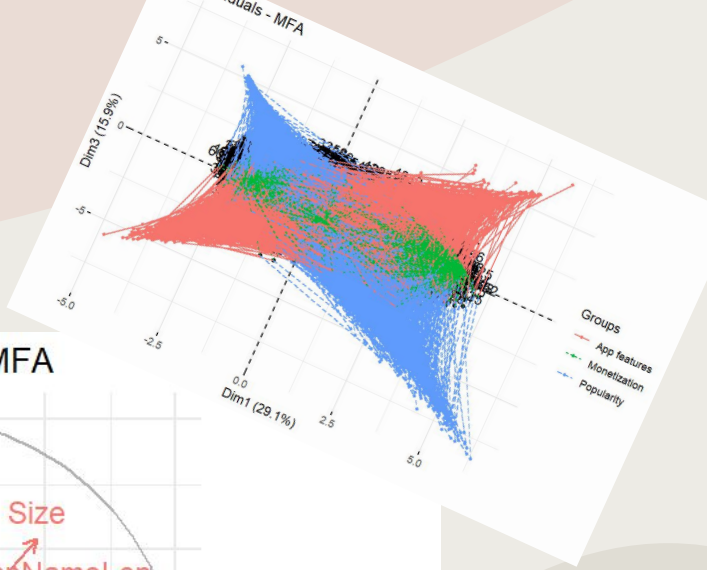
Dimensions 1 & 2

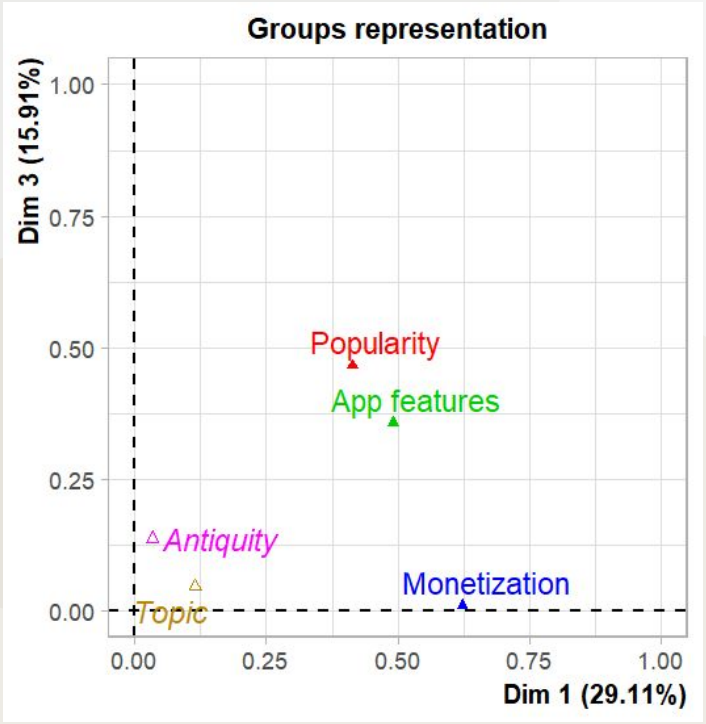


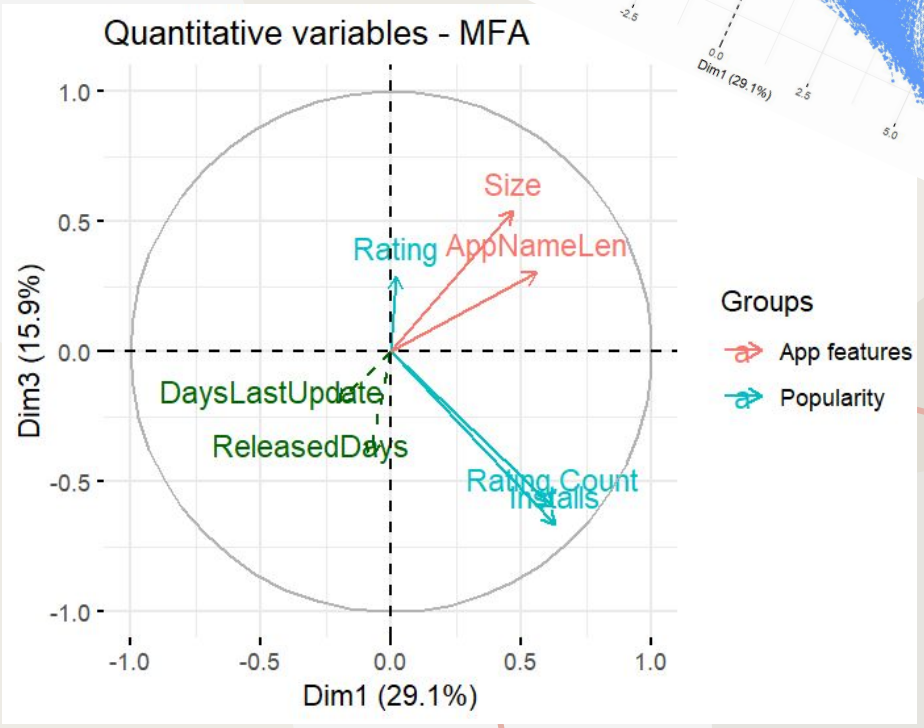




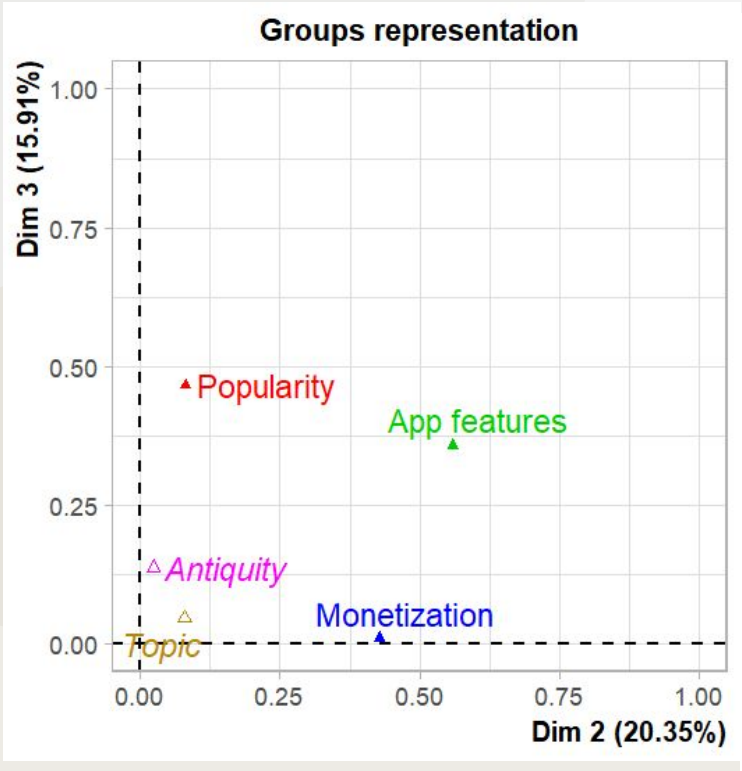
Dimensions 1 & 3

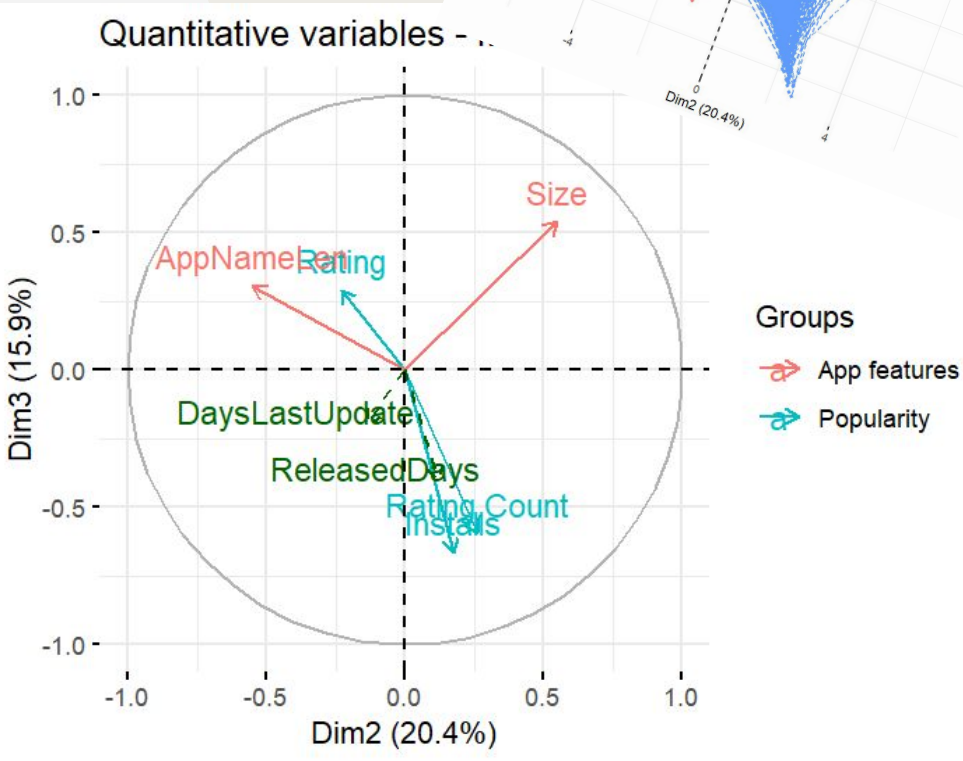






Dimensions 2 & 3

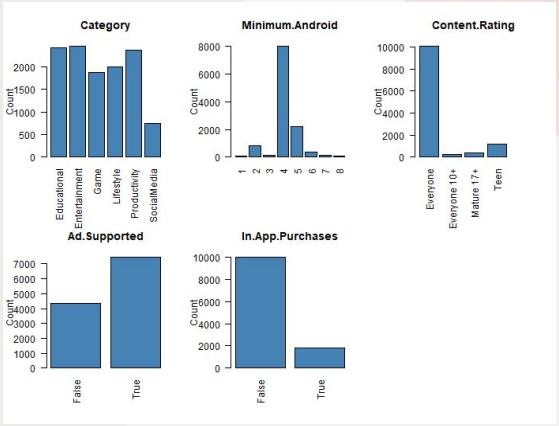


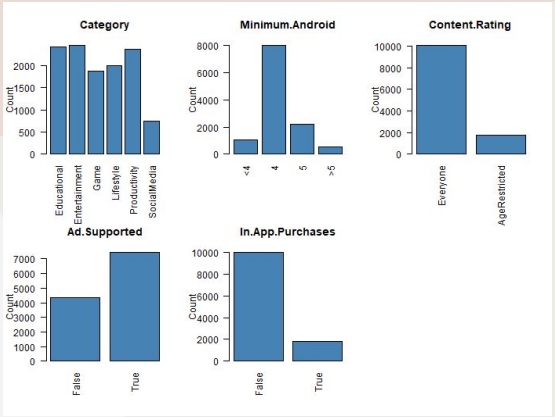


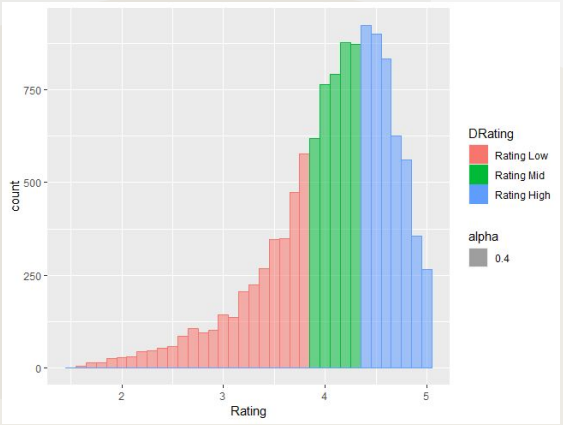
09

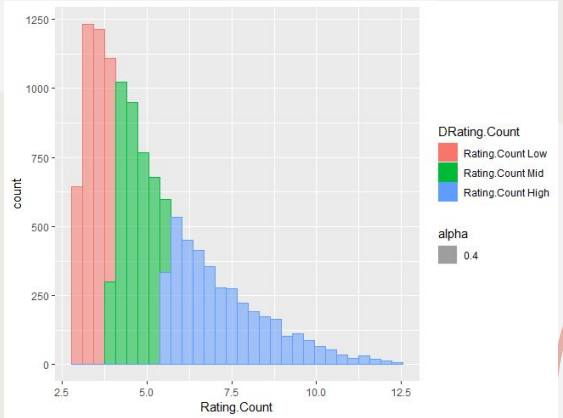
Association Rules

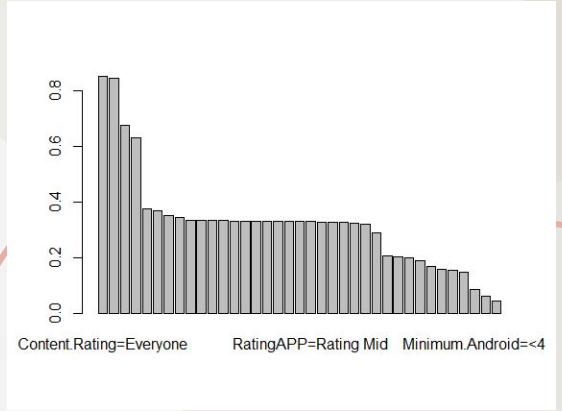
Association Rules











Association Rules

Content Rating = Everyone &
In App Purchases = False &
Installs = Low

Rating = High

Content Rating = Everyone &
Installs = Low

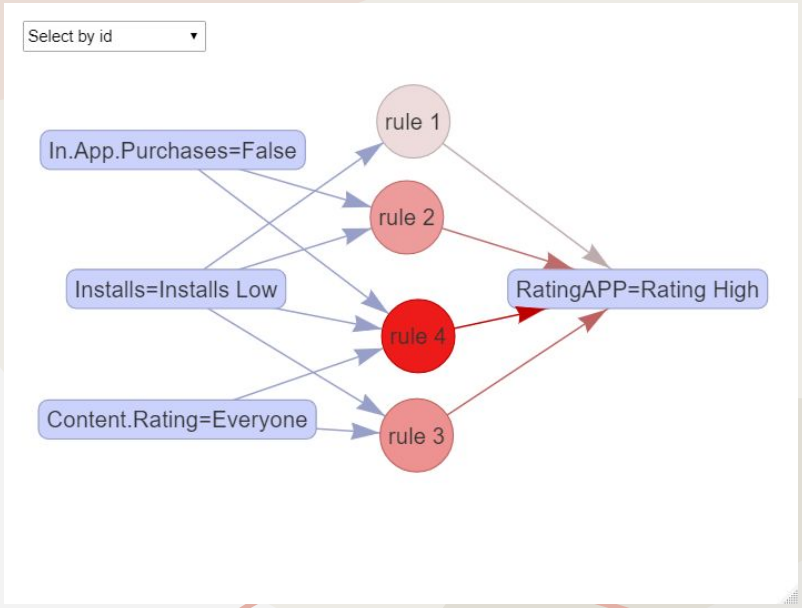
Rating = High

In App Purchases = False &
Installs = Low

Rating = High

Content Rating = Everyone &
In App Purchases = False &
Installs = Low

Rating = High



0.14 < **Support** < 0.19

0.54 < **Confidence** < 0.56

1.4 < **Lift** < 1.5

Association Rules

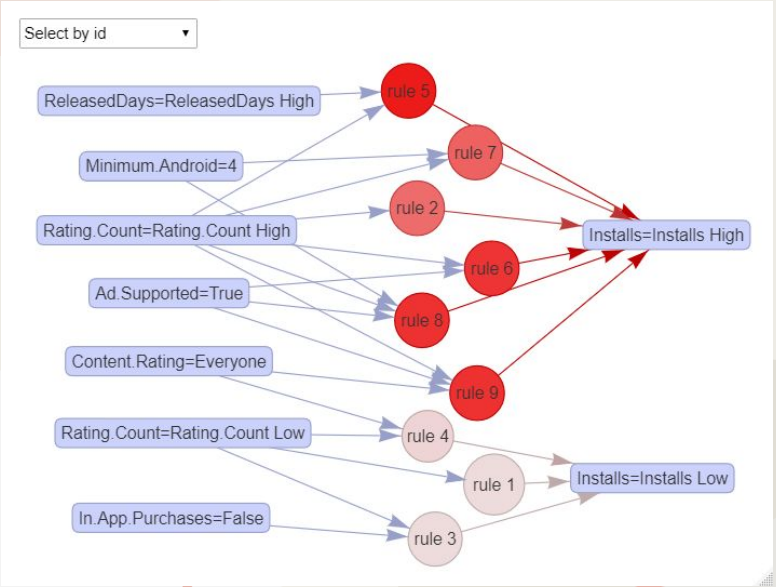
9 Rules

Rating Count = High &
Release Days = High → Installs = High

Rating Count = High &
Content Rating = Everyone → Installs = High

Rating Count = High &
Content Rating = Everyone &
Ad Supported = True → Installs = High

... ...



0.13 < **Support** < 0.28

0.72 < **Confidence** < 0.88

2.18 < **Lift** < 2.62

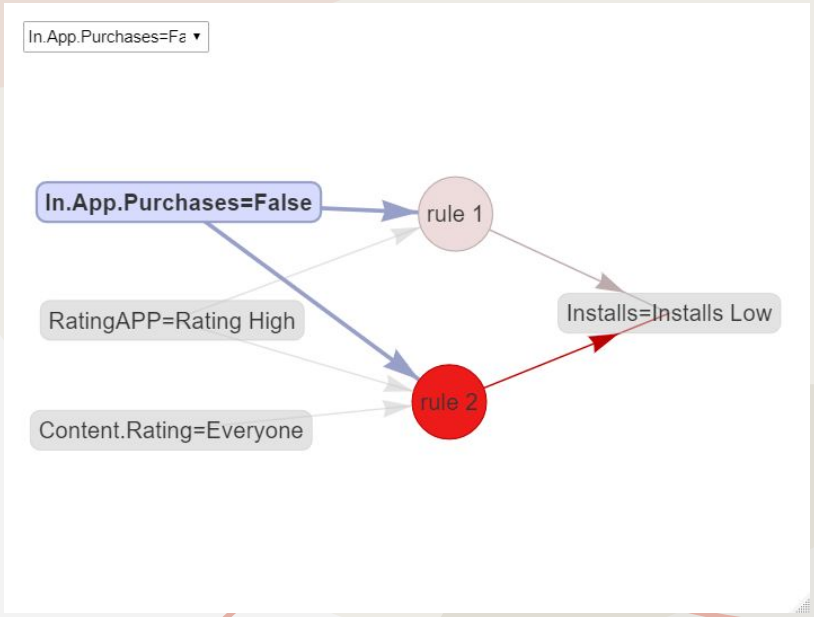
Association Rules

Content Rating = Everyone &
In App Purchases = False &
Rating = High

Installs = Low

In App Purchases = False &
Rating = High

Installs = Low



0.14 < **Support** < 0.17

0.51 < **Confidence** < 0.53

1.53 < **Lift** < 1.57

Thank you for your attention

Any questions?