UNSUPERVISED METHODS TO GROUP USERS’ CONSUMPTION BEHAVIOUR TO ENHANCE PERSONALIZE SERVICE DEGRADATION POLICIES

PROJECT MODULE 1

ROBERTO CASALUCE – MACIEJ ZUZIAK

EXPLORATORY DATA ANALYSIS AND FEATURE ENGINEERING

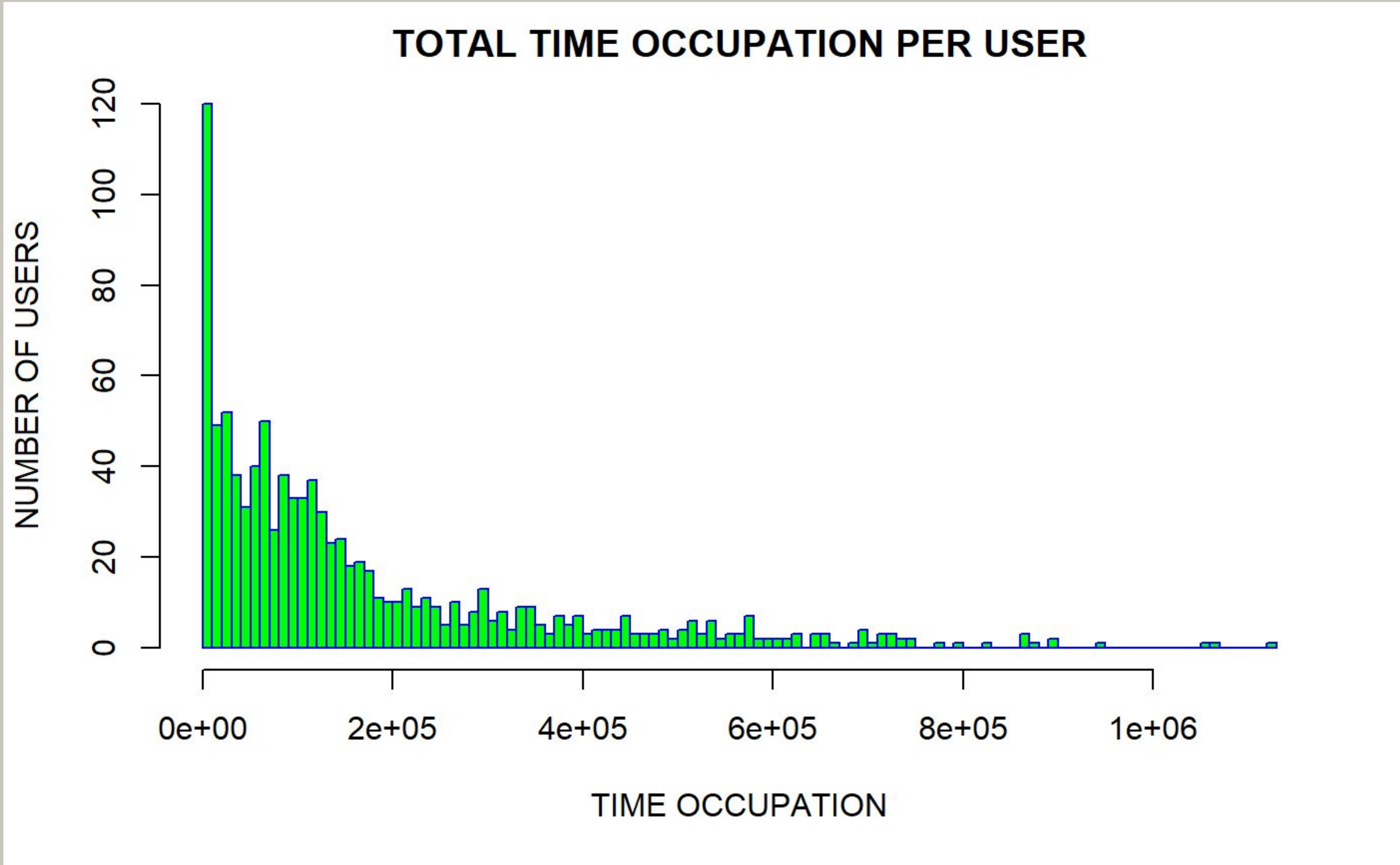
OVERVIEW

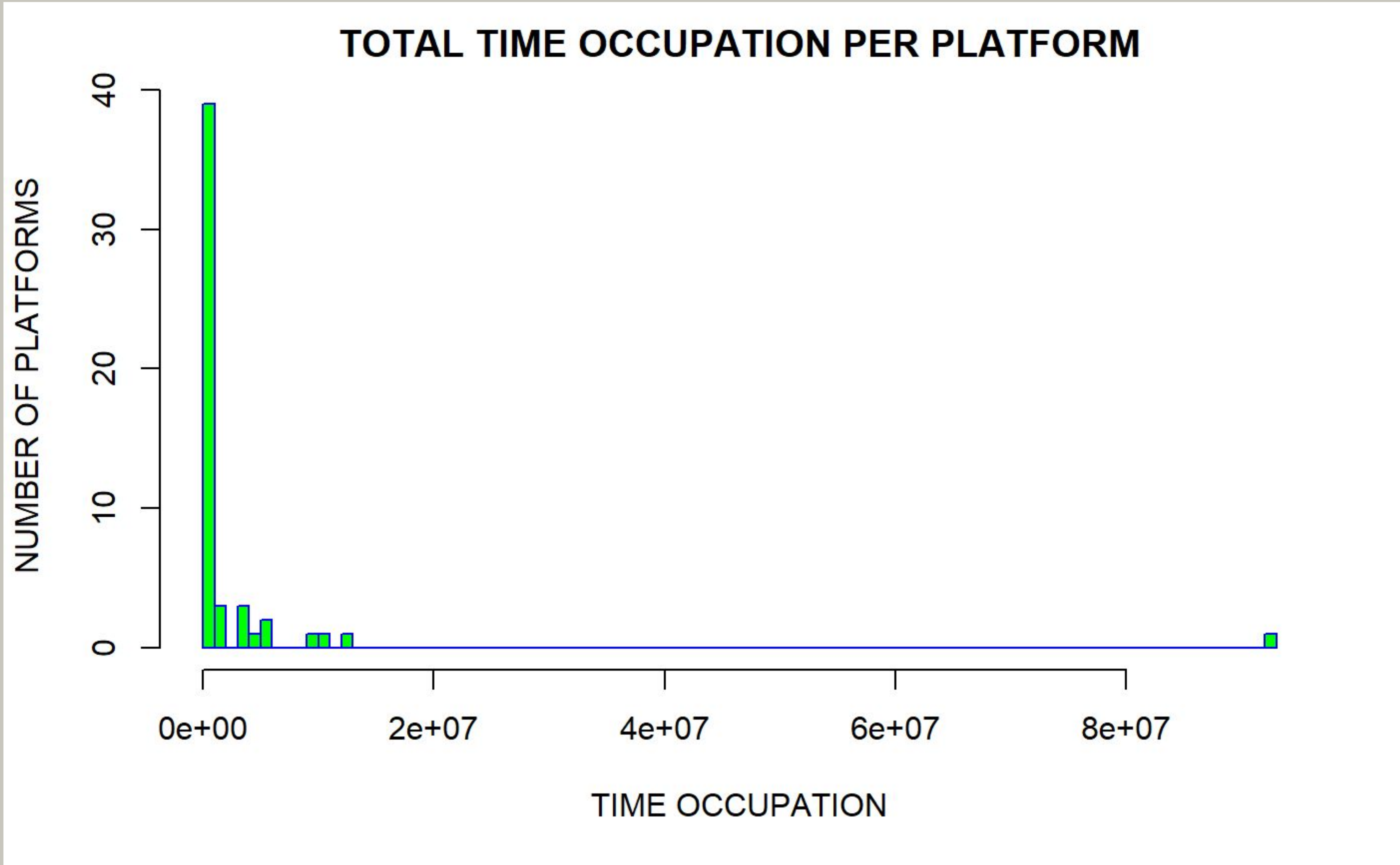
ONE DATASET CONTAINING 1249 STATISTICAL UNITS

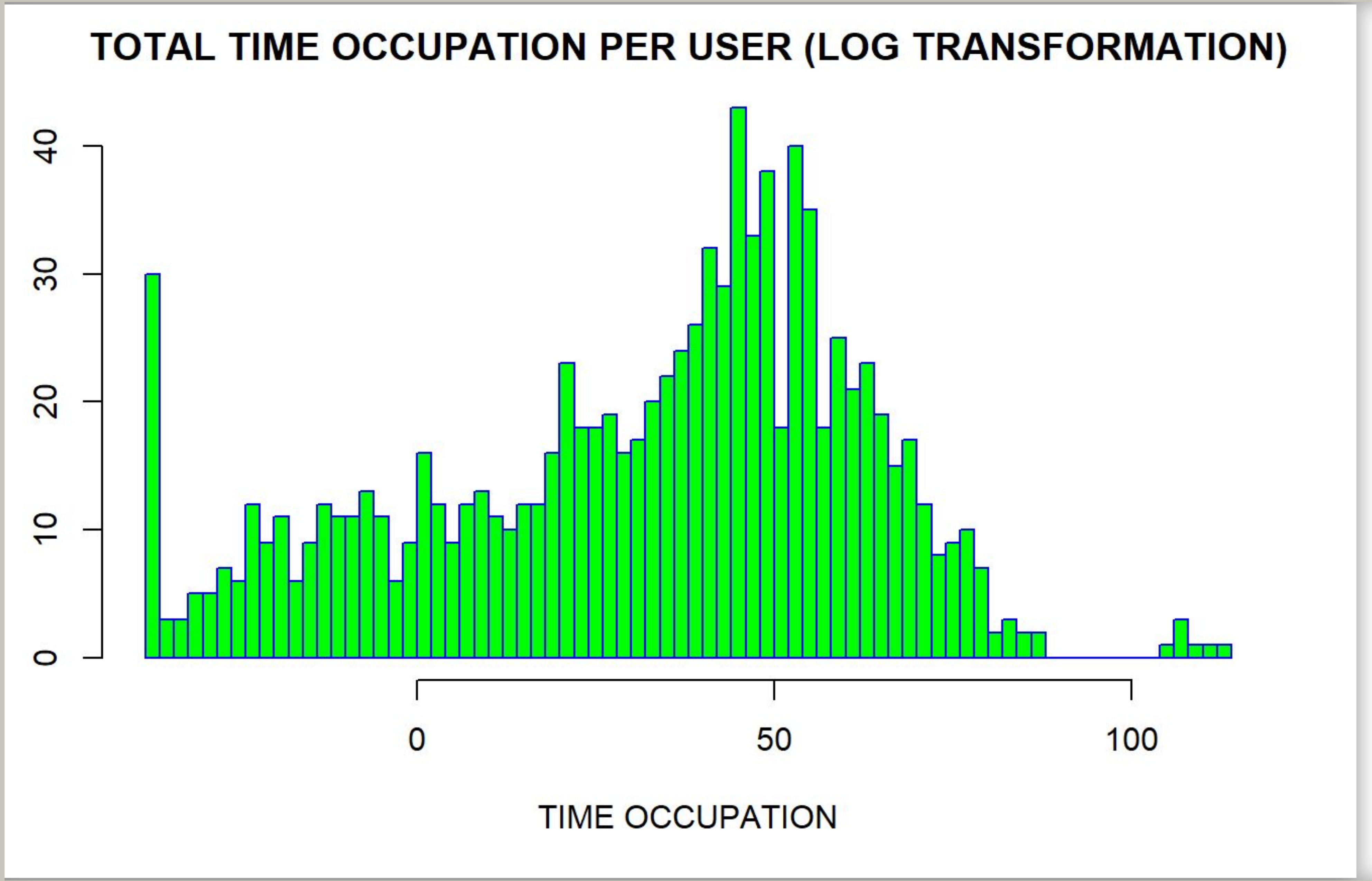
114 DIFFERENT FEATURES (ONLINE PLATFORMS AND MOBILE APPLICATIONS)

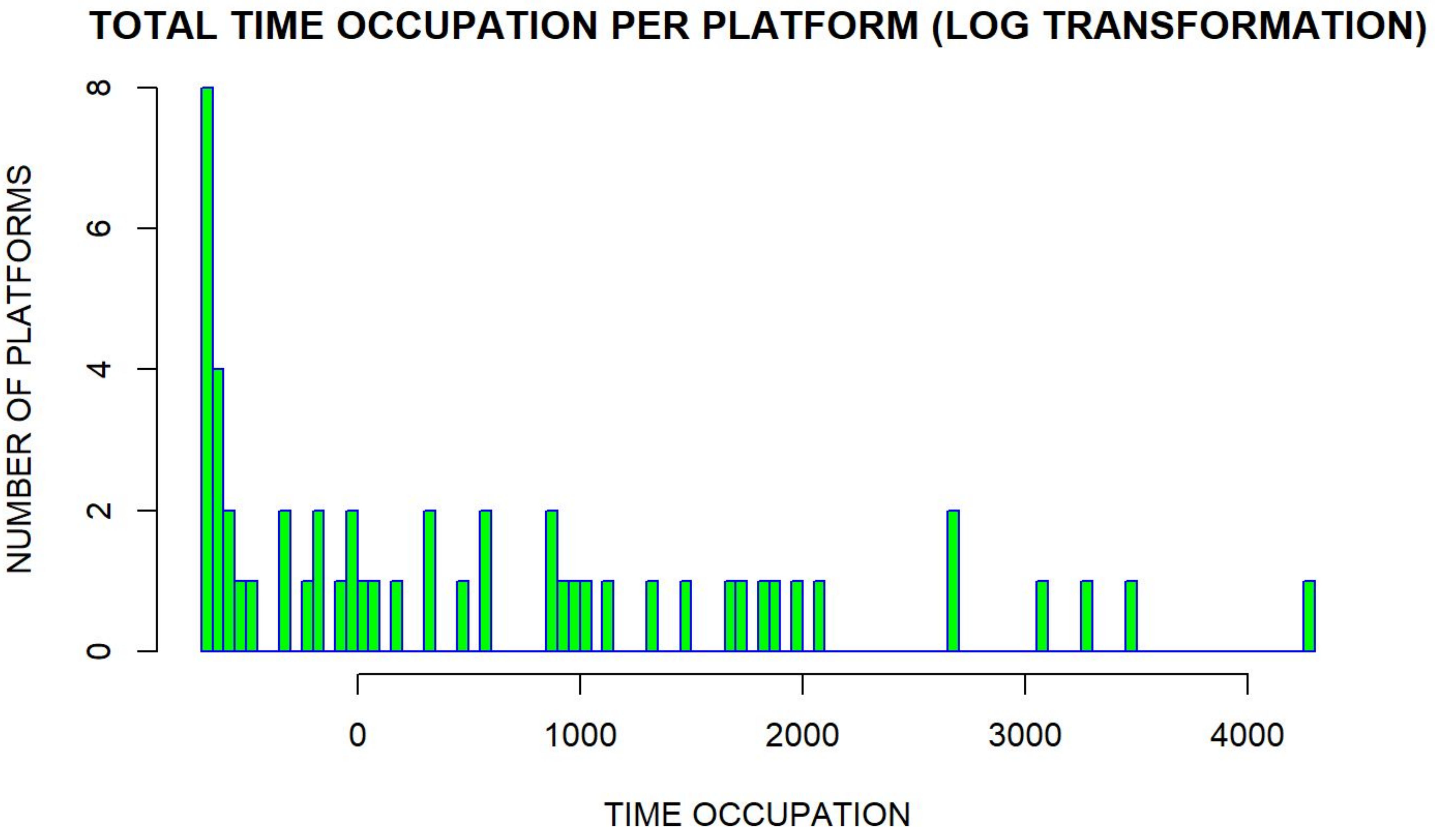
TIME EXPRESSED IN SECOND, DATA CONSUMPTION IN BYTES

TASK IS TO IDENTIFY THE EXISTENCE OF CONSUMER GROUPS


TOTAL TIME OCCUPATION PER USER
NUMBER OF USERS
0
20
40
60
80
100
120
0e+00
2e+05
4e+05
6e+05
8e+05
1e+06
TIME OCCUPATION

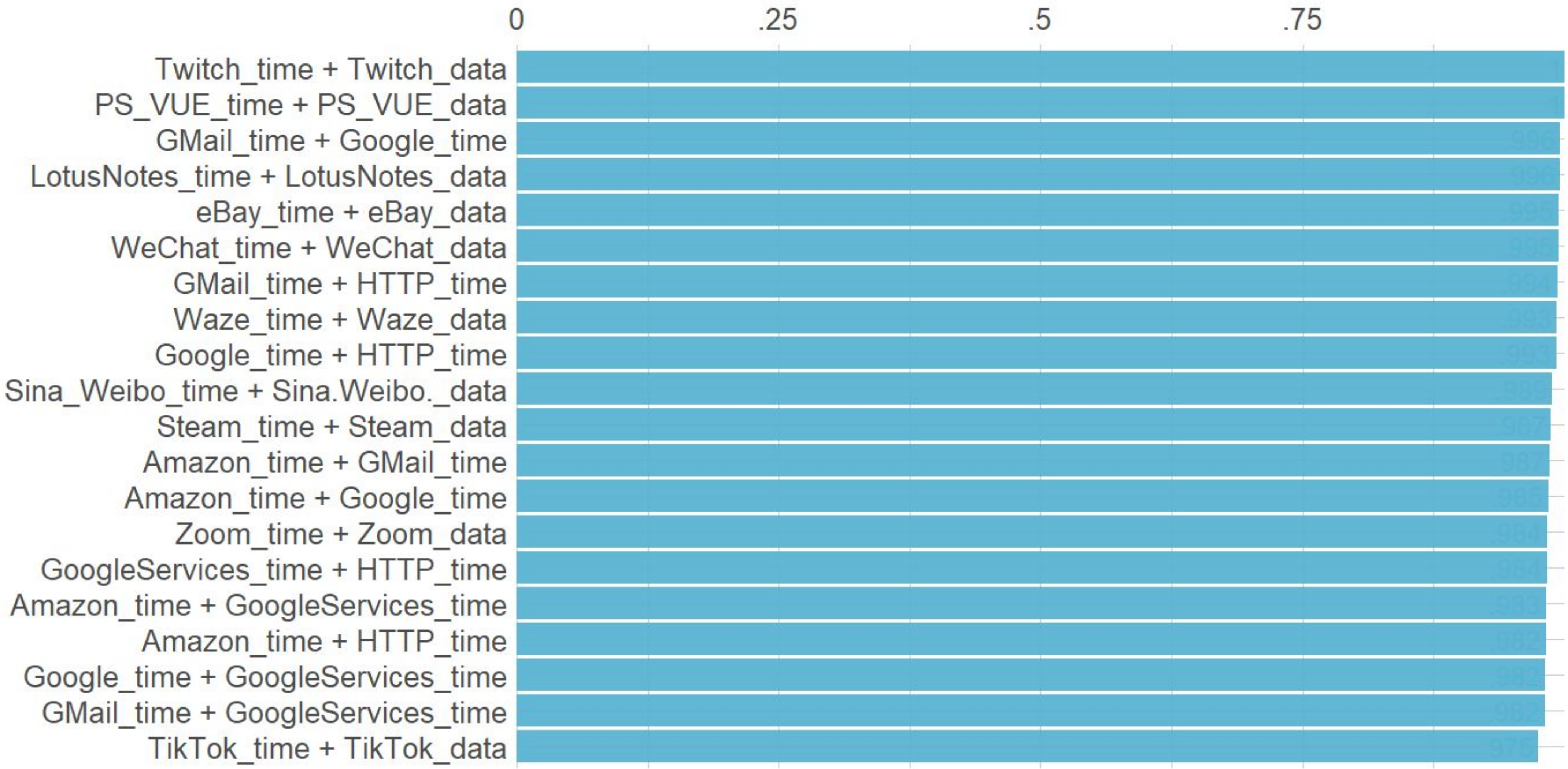

TOTAL TIME OCCUPATION PER PLATFORM
NUMBER OF PLATFORMS
0
10
20
30
40
0e+00
2e+07
4e+07
6e+07
8e+07
TIME OCCUPATION


TOTAL TIME OCCUPATION PER USER (LOG TRANSFORMATION)
0
10
20
30
40
0
50
100
TIME OCCUPATION


TOTAL TIME OCCUPATION PER PLATFORM (LOG TRANSFORMATION)
NUMBER OF PLATFORMS
0
2
4
6
8
0
1000
2000
3000
4000
TIME OCCUPATION

Ranked Cross-Correlations

20 most relevant



Correlations with p-value < 0.05

FEATURE ENGINEERING OVERVIEW

WE HAVE…

01

DELETED FEATURES RELATED TO DATA USAGE

02

DELETED FEATURES THAT CONTAINED LESS THAN TWO USERS' ENTRIES

03

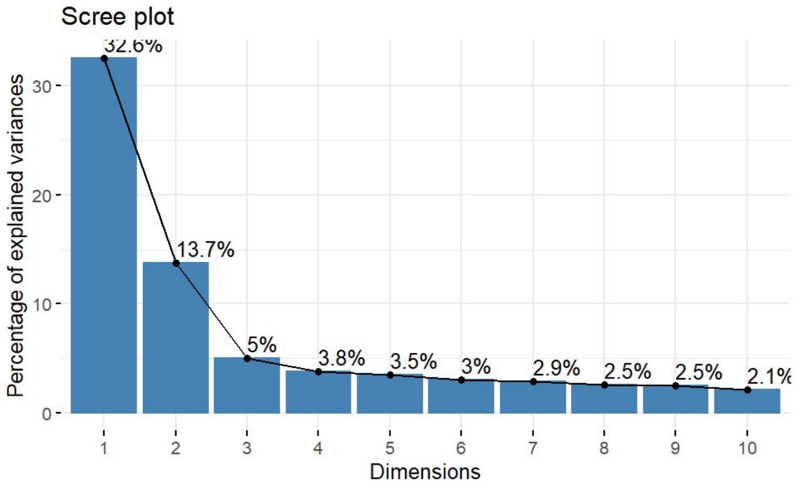
DELETED DUPLICATED ENTRIES

04

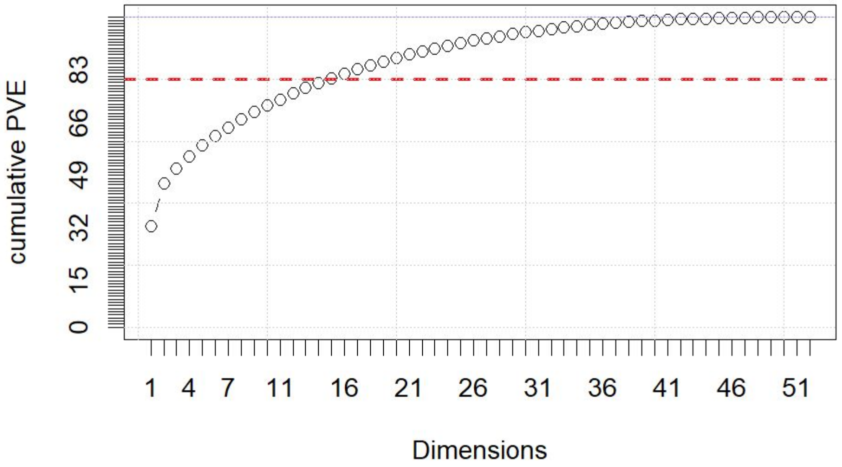
PERFORMED LOG TRANSFORMATION DESCRIBED BEFORE

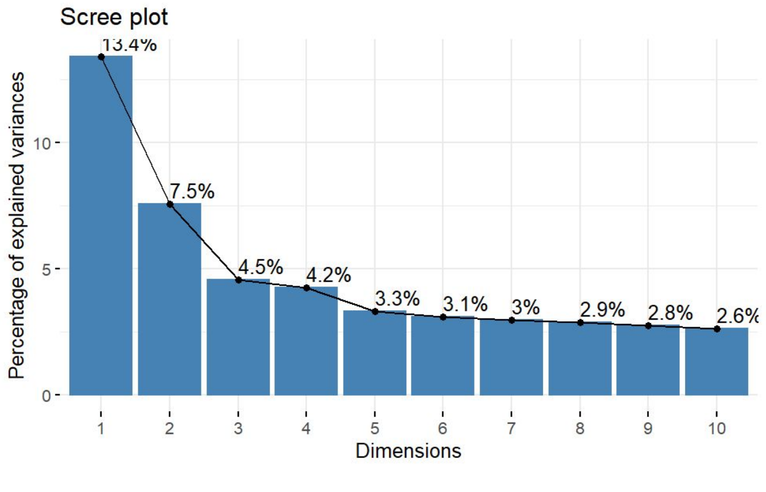
DATASET VISUALIZATION (PCA)

PCA

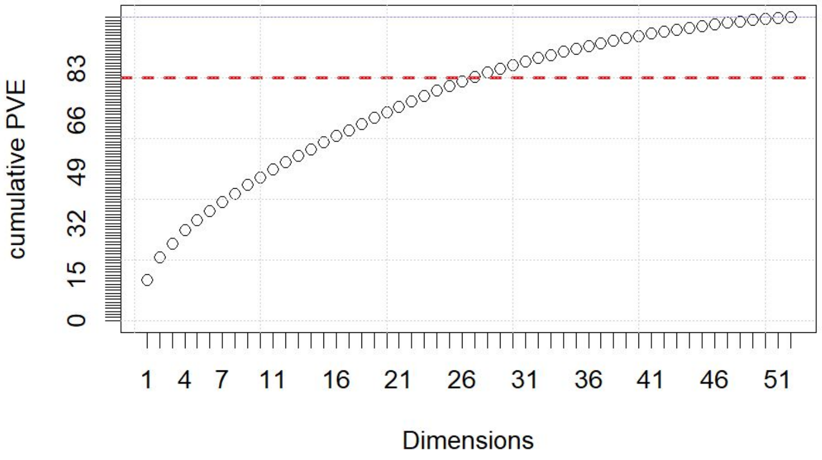


Log transformed data

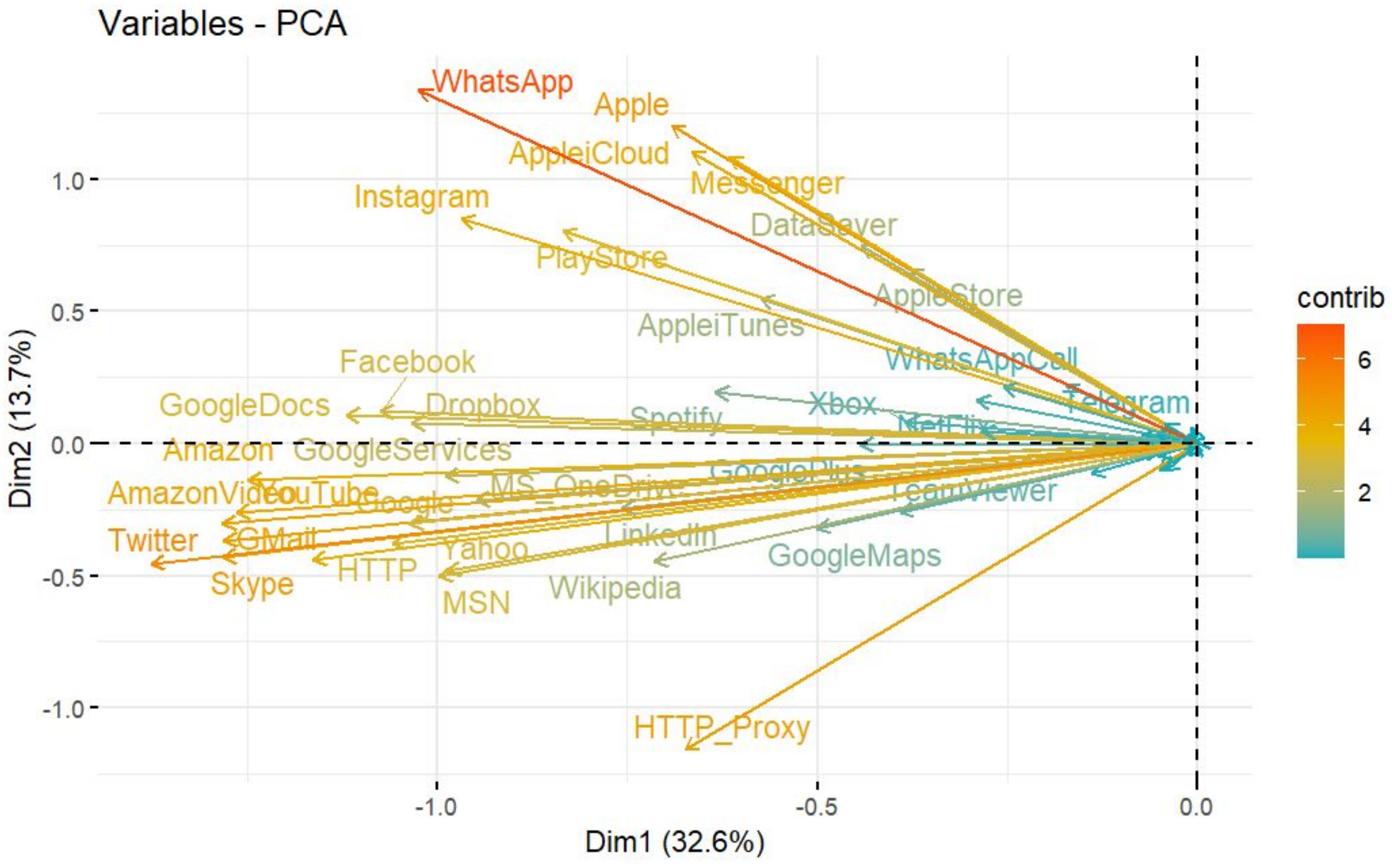


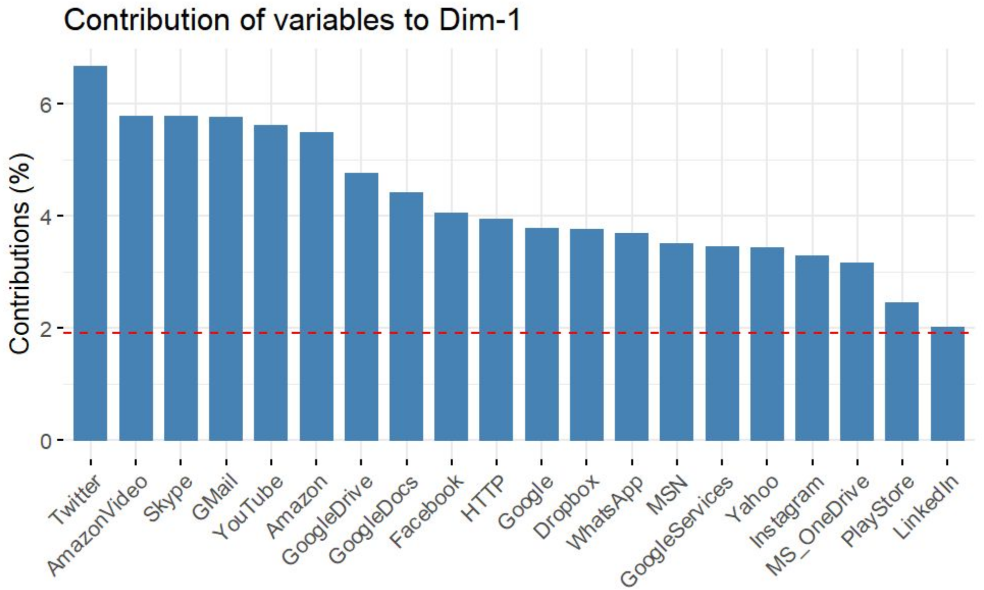


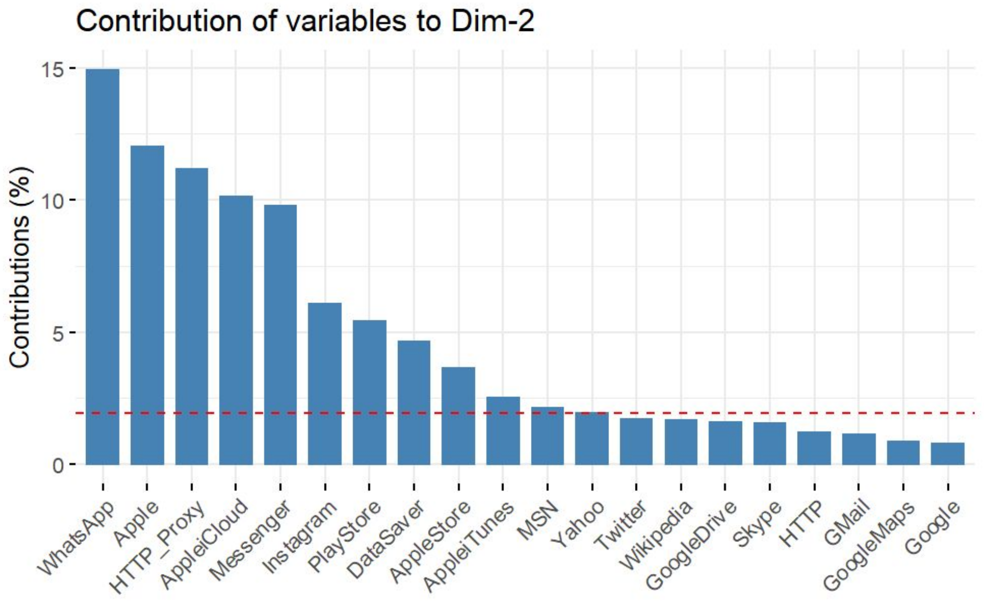
Scaled data



PCA


Variables - PCA
Dim1 (32.6%)
Dim2 (13.7%)
-1.0
-0.5
0.0
0.5
1.0
contrib
6
4
2
WhatsApp
Apple
AppleiCloud
Messenger
Instagram
DataSaver
PlayStore
AppleStore
AppleiTunes
WhatsAppCall
Facebook
GoogleDocs
Dropbox
Spotify
Xbox
Telegram
NetFlix
Amazon
GoogleServices
GooglePlus
MS_OneDrive
AmazonVideo
YouTube
Google
TeamViewer
Twitter
GMail
LinkedIn
Yahoo
GoogleMaps
HTTP
Skype
MSN
Wikipedia
HTTP_Proxy

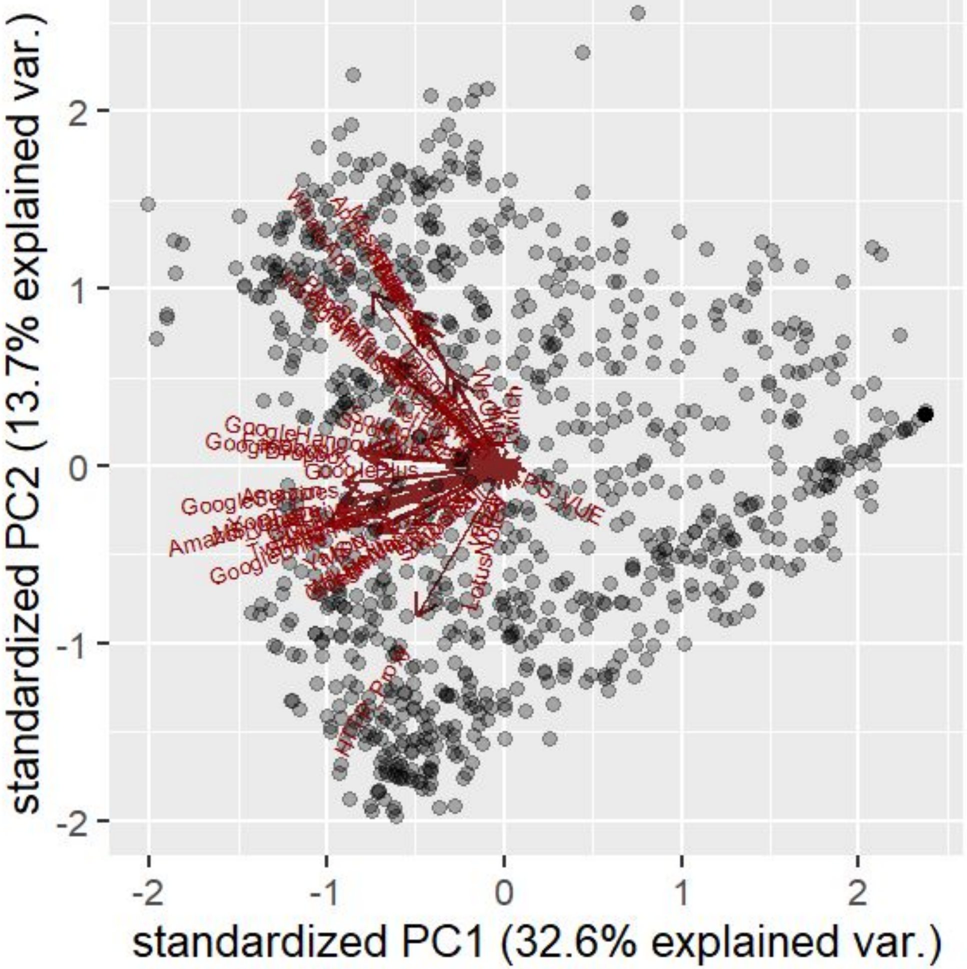

Contribution of variables to Dim-1
Contributions (%)
0
2
4
6
Twitter
AmazonVideo
Skype
GMail
YouTube
Amazon
GoogleDrive
GoogleDocs
Facebook
HTTP
Google
Dropbox
WhatsApp
MSN
GoogleServices
Yahoo
Instagram
MS_OneDrive
PlayStore
LinkedIn


Contribution of variables to Dim-2
Contributions (%)
0
5
10
15
WhatsApp
Apple
HTTP_Proxy
AppleiCloud
Messenger
Instagram
PlayStore
DataSaver
AppleStore
AppleiTunes
MSN
Yahoo
Twitter
Wikipedia
GoogleDrive
Skype
HTTP
GMail
GoogleMaps
Google

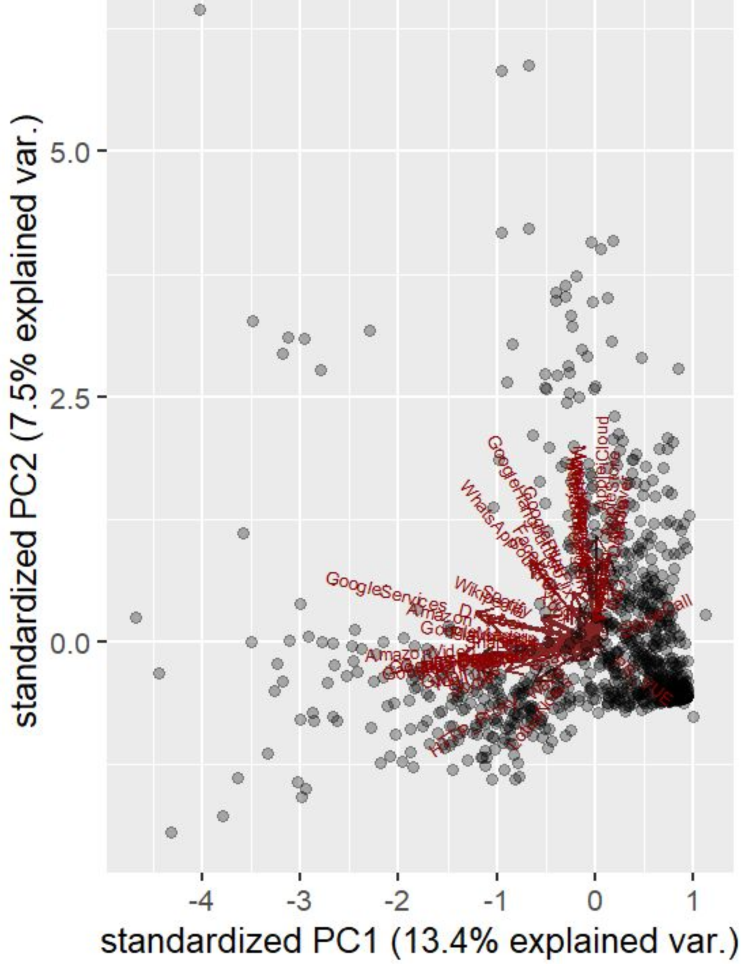
PCA

Biplots

Log transformed data



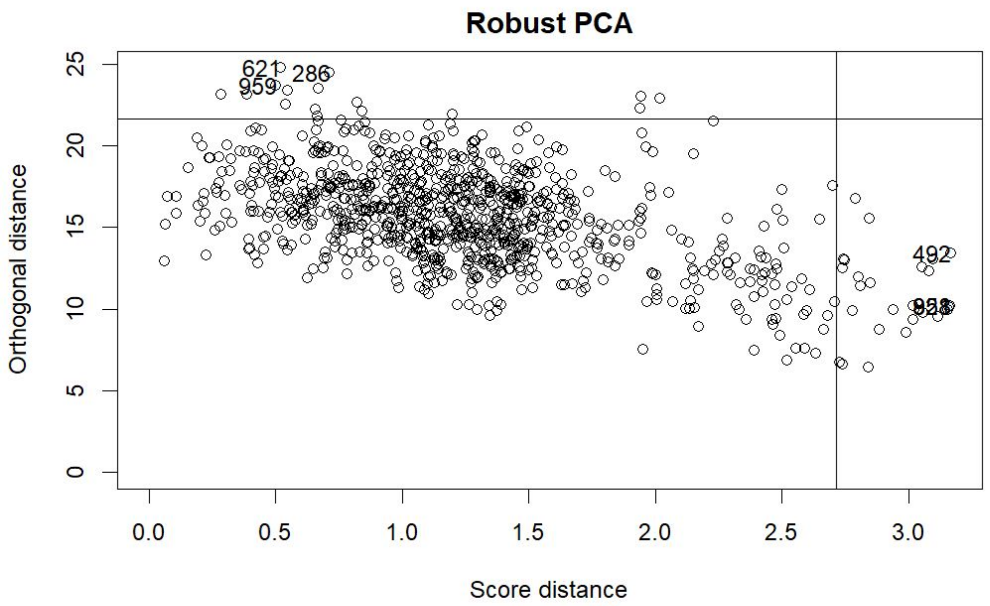
Scaled data



Robust PCA

Flags the outliers

First Two components



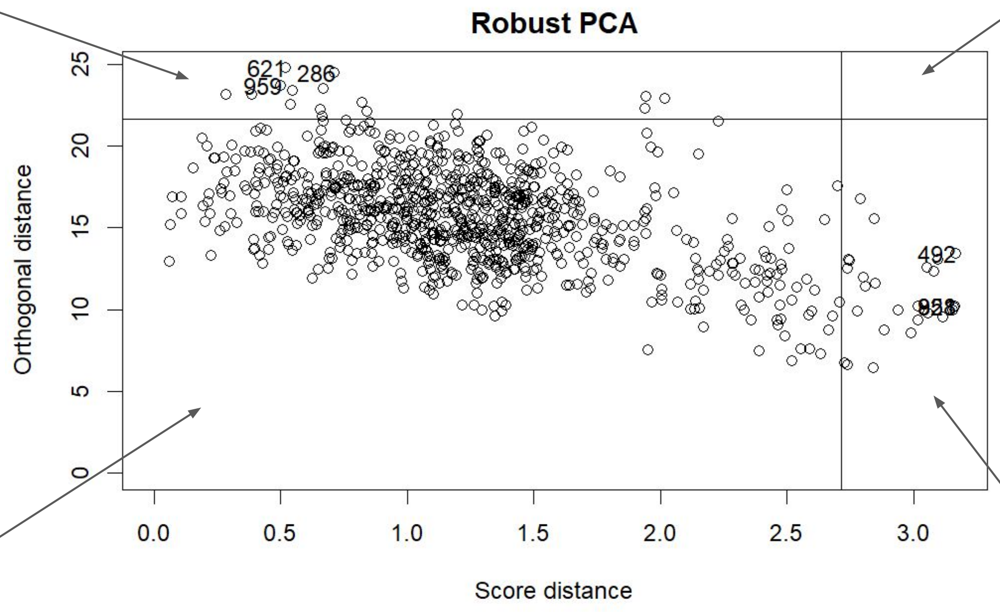
Robust PCA

Flags the outliers

Orthogonal outliners

Bad leverage

First Two components

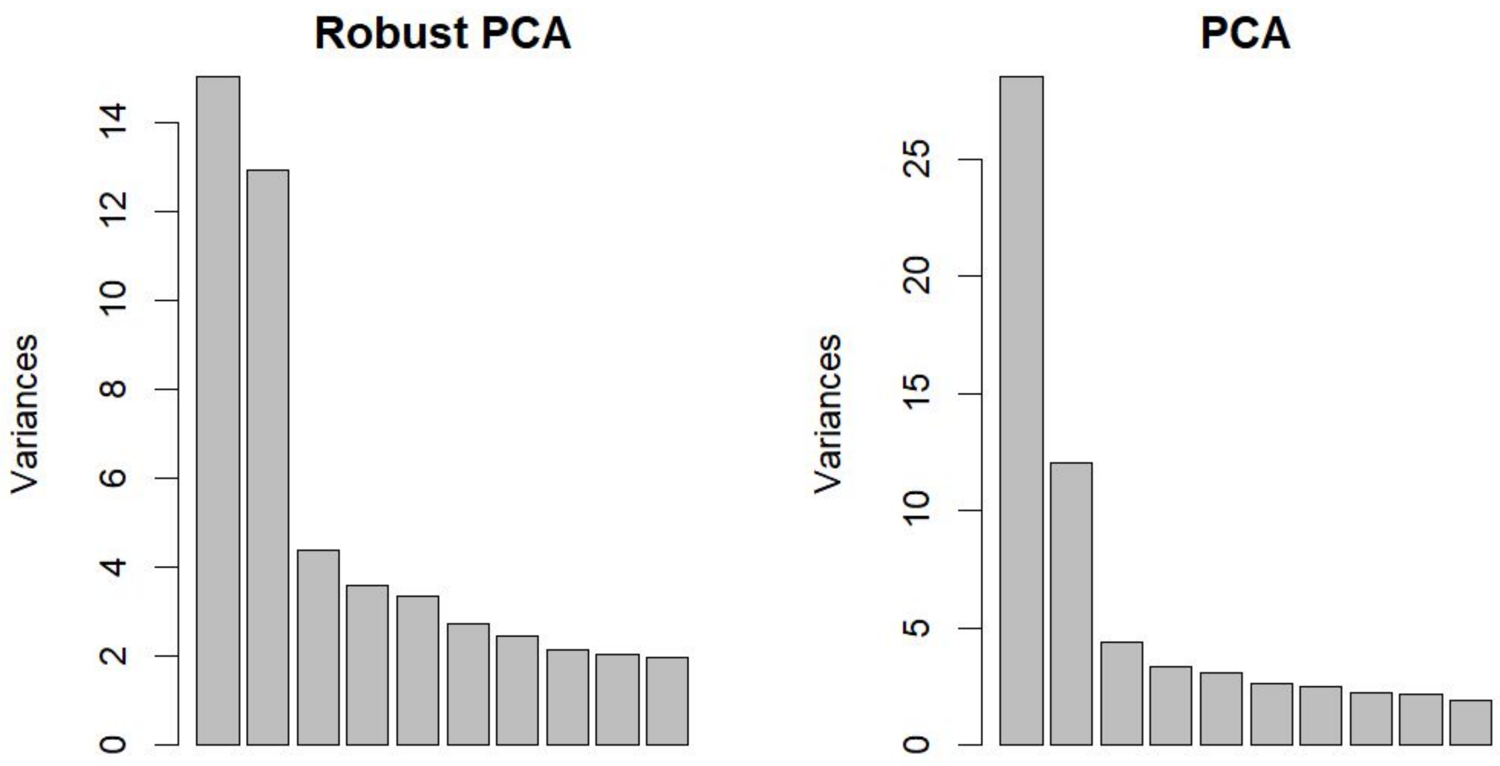


Regular points

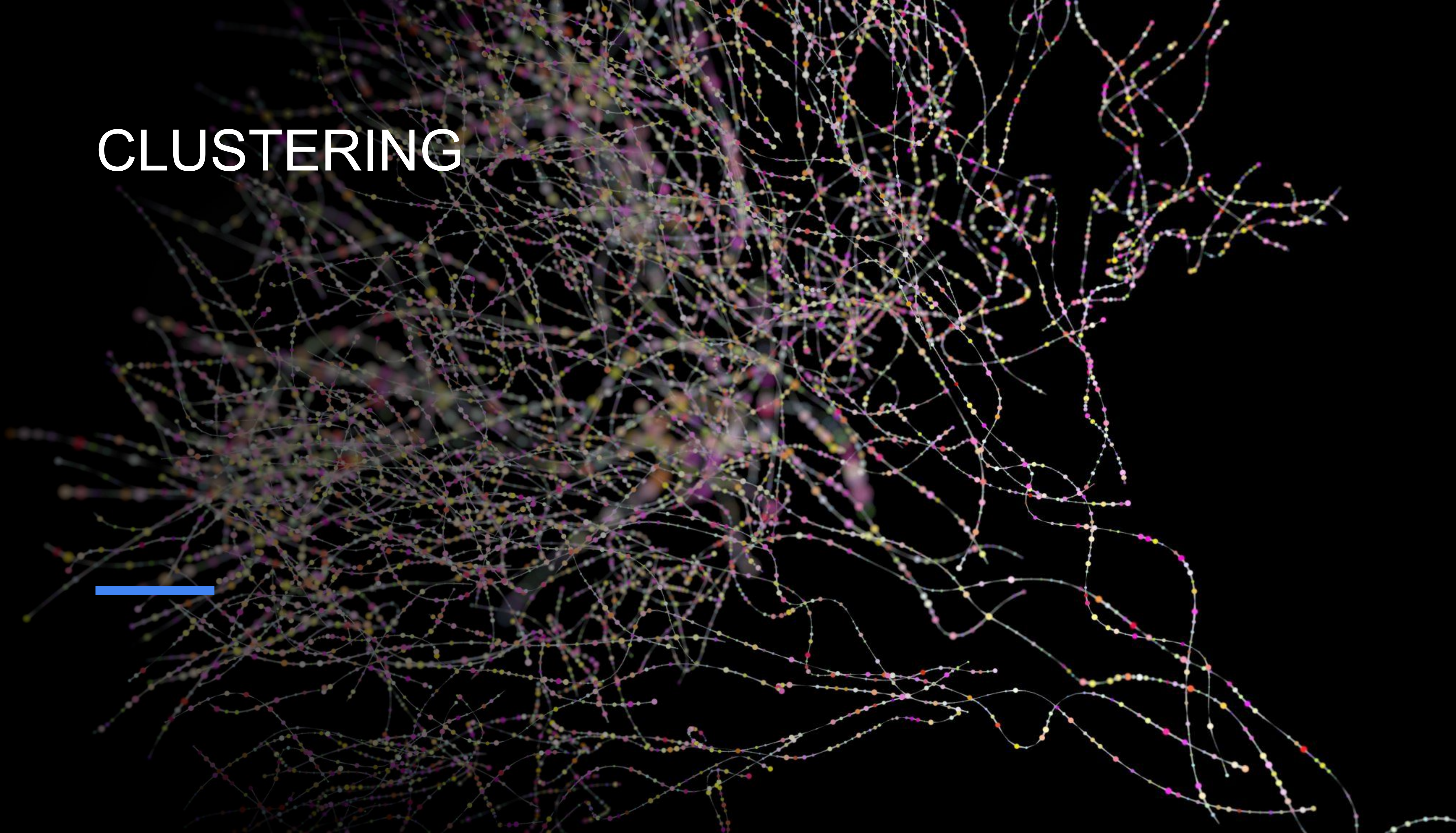
Good leverage

(Chen et al., 2020)

Variances Robust and Classic PCA



CLUSTERING

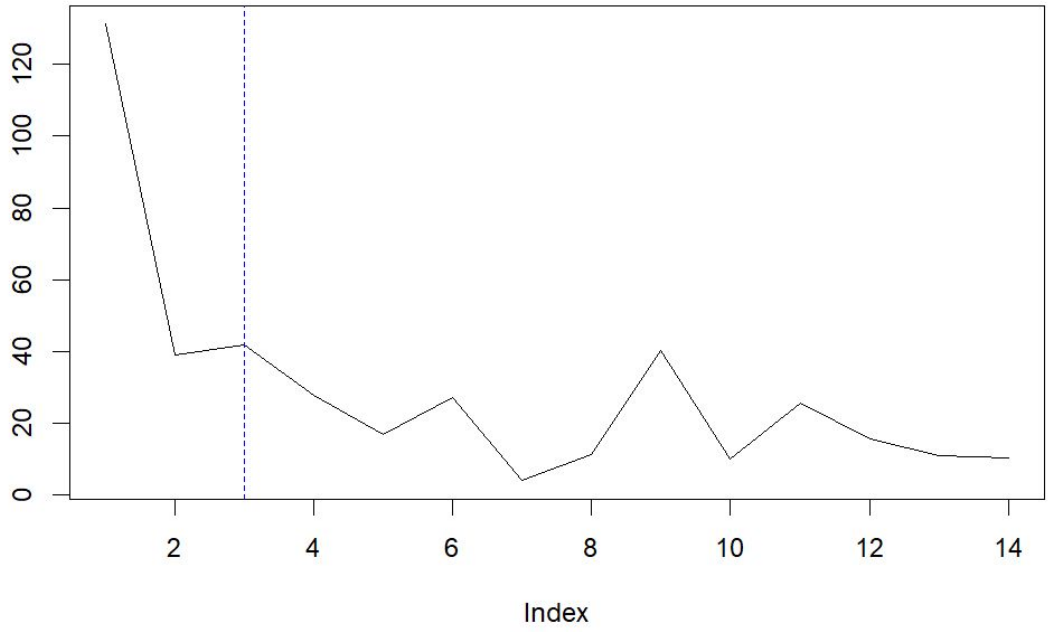


Clustering method selection

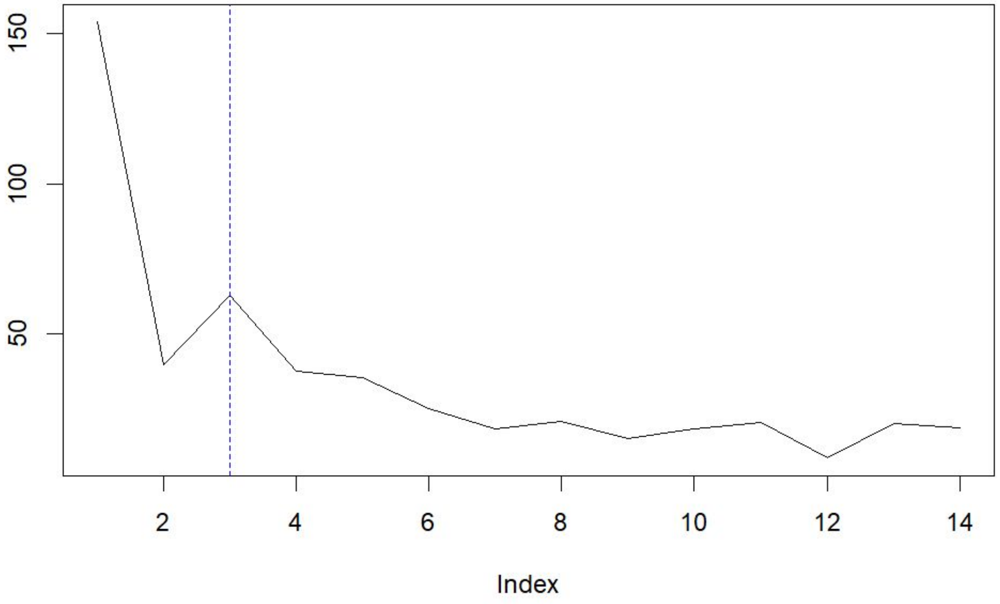
From a previous work (Rojas et al., 2020):
3/4 clusters - **Low**, **Medium, High and Very High**

Hartigan Index

Agglomerative Hierarchical clustering method



Kmeans clustering method



Evaluating Clustering methods (Brock et al., 2008)

Cluster **stability** measure:

- The average proportion of non-overlap (APN)
- The average distance (AD)
- The average distance between means (ADM)
- The figure of merit (FOM)

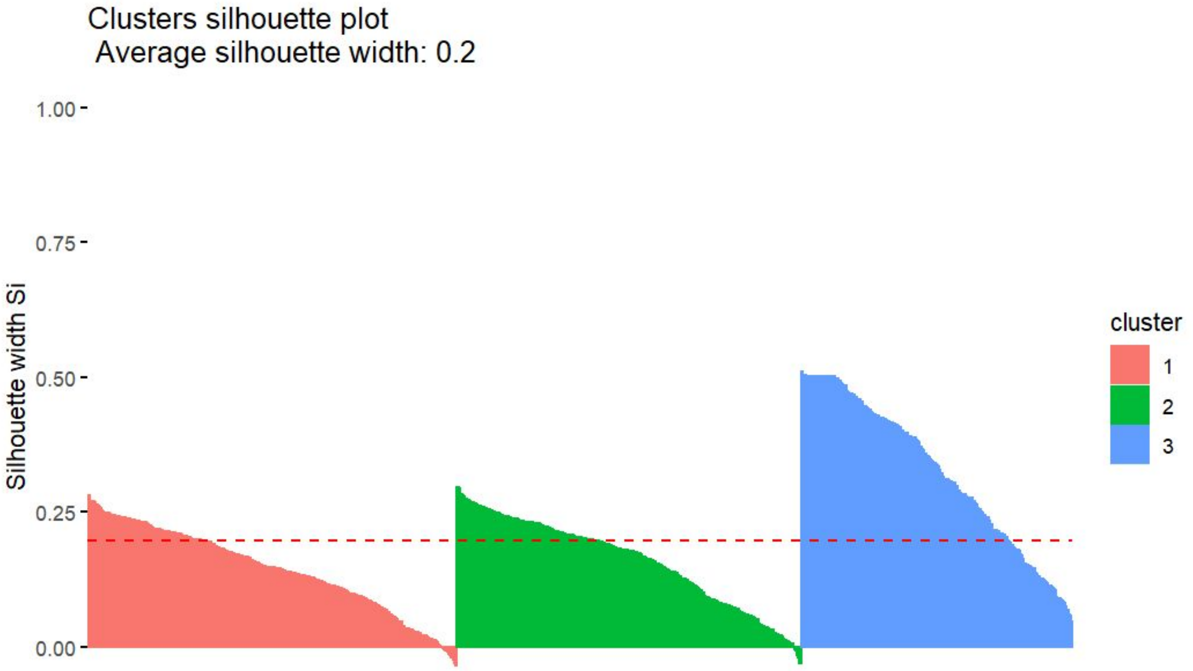
APN	0.009	kmeans	3
AD	10.2	kmeans	4
ADM	0.095	kmeans	3
FOM	0.93	kmeans	4

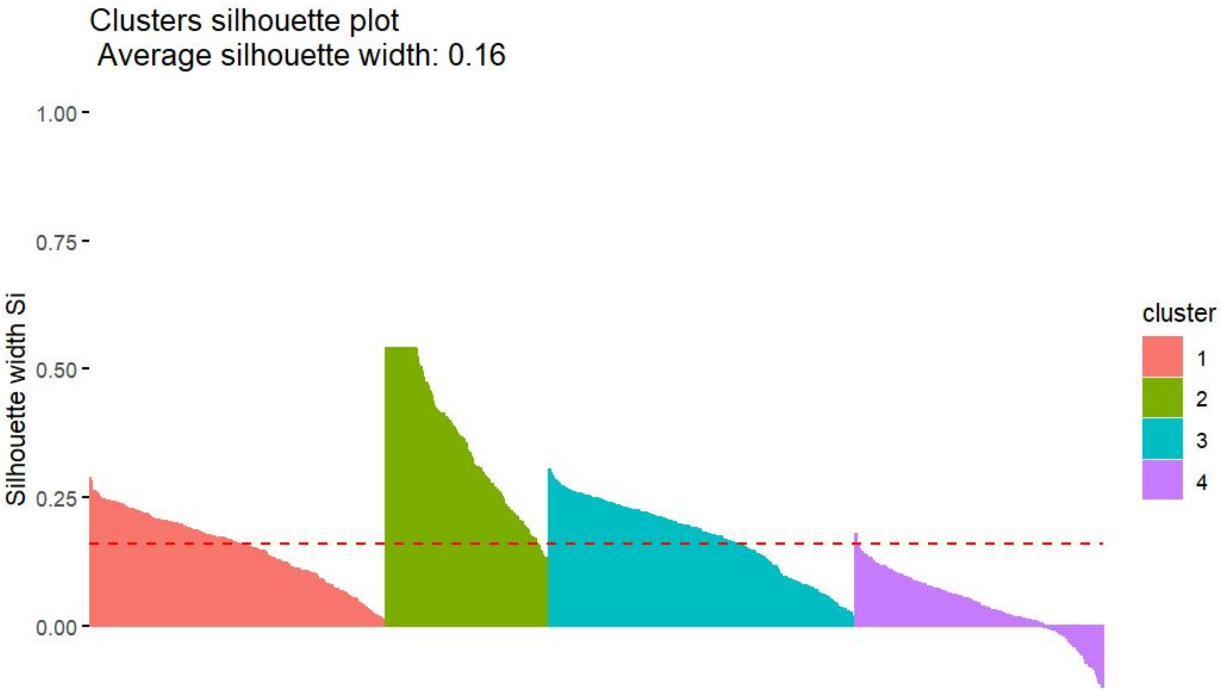
Connectivity	139.75	hierarchical	3
Dunn	0.26	hierarchical	3
Silhouette	0.20	kmeans	3

Internal measures for cluster validation

Evaluating a Clustering Solution

Silhouette widths kmeans

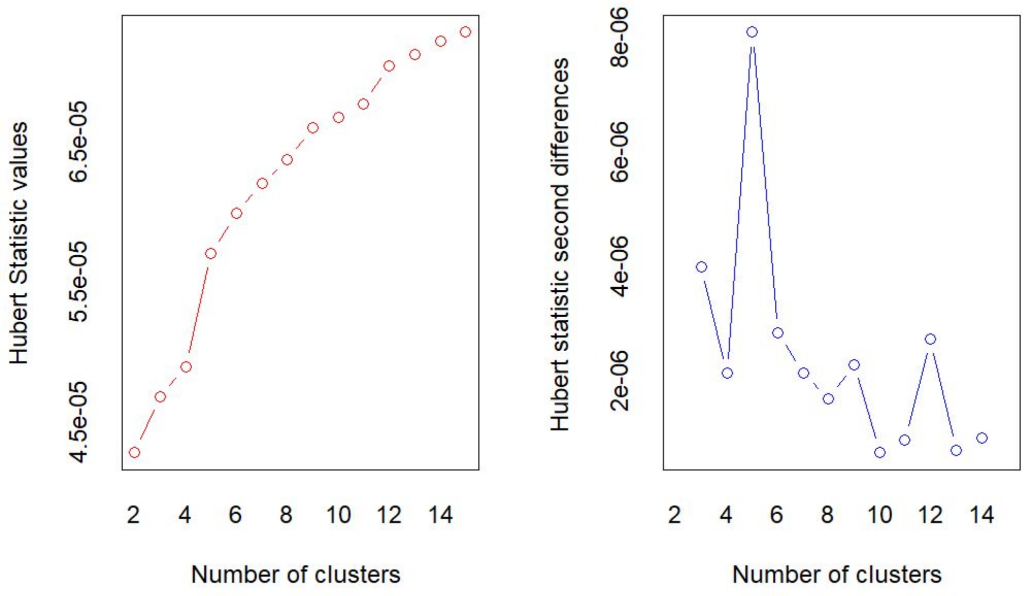




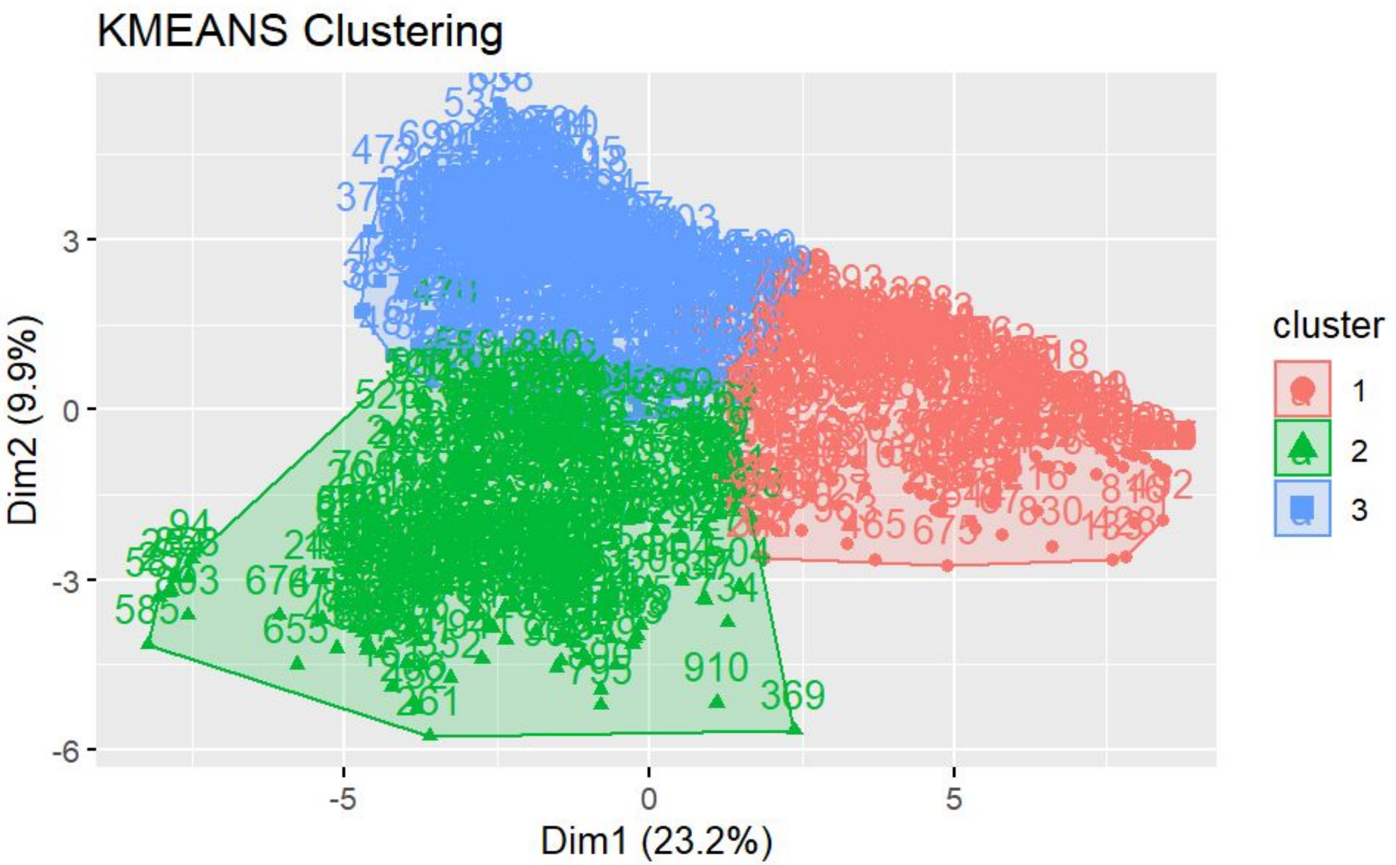
Evaluating a Clustering Solution

Hubert index kmeans

According to the majority rule, the best number of clusters is 3
* 11 proposed 3 as the best number of clusters
* 1 proposed 4 as the best number of clusters
* 9 proposed 2 as the best number of clusters



Kmeans 3 clusters



Size Clusters		
1	2	3
266	343	364

Reference

Borg, I. and Groenen, P. (1997) Modern Multidimensional Scaling. Theory and Applications. Springer.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The New S Language. Wadsworth & Brooks/Cole.

Brock, G., Pihur, V., Datta, S. and Datta, S. (2008) clValid: An R Package for Cluster Validation Journal of Statistical Software 25(4)

Charrad M., Ghazzali N., Boiteau V., Niknafs A. (2014). "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set.", "Journal of Statistical Software, 61(6), 1-36.", "URL http://www.jstatsoft.org/v61/i06/".

Chen, X., Zhang, B., Wang, T., Bonni, A., & Zhao, G. (2020). Robust principal component analysis for accurate outlier sample detection in RNA-Seq data. *BMC bioinformatics*, *21*(1), 1-20.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency vs interpretability of classifications. Biometrics, 21, 768–769.

Reference

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. Applied Statistics, 28, 100–108. doi: 10.2307/2346830.

Lloyd, S. P. (1957, 1982). Least squares quantization in PCM. Technical Note, Bell Laboratories. Published in 1982 in IEEE Transactions on Information Theory, 28, 128–137.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, eds L. M. Le Cam & J. Neyman, 1, pp. 281–297. Berkeley, CA: University of California Press.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979) Multivariate Analysis. Academic Press.

Metcalf, L., & Casey, W. (2016). Cybersecurity and applied mathematics. Syngress.

Rojas, J. S., Pekar, A., Rendón, Á., & Corrales, J. C. (2020). Smart user consumption profiling: Incremental learning-based OTT service degradation. IEEE access, 8, 207426-207442.