User Analytics in the Telecommunication Industry

OUTLINE

- Introduction
- Non Graphical Univariate Analysis
- Graphical Univariate Analysis
- Graphical Bivariate analysis
- Correlation analysis
- Dimensionality reduction
- Engagement metrics analysis
- Experience metrics analysis
- Satisfaction metrics
- Database
- Conclusion and recommendation

INTRODUCTION

Description of Datasets

- The metrics used in this study are grouped based on their implication or uses
- The duration of the session , the session frequency, total traffic in a session can be used to track the engagement of a user
- Experience metrics focuses on the needs and requirements of the customer, this can be evaluated from the network parameters
- Satisfaction metrics are derived metrics that explains how content a user is towards the services

Fields	Description
Dur. (ms)	Total Duration of the xDR (in ms)
MSISDN/Number	MS International PSTN/ISDN Number of mobile - customer number
Avg RTT DL (ms)	Average Round Trip Time measurement Downlink direction (msecond)
Avg RTT UL (ms)	Average Round Trip Time measurement Uplink direction (msecond)
Avg Bearer TP DL (kbps)	Average Bearer Throughput for Downlink (kbps) - based on BDR duration
Avg Bearer TP UL (kbps)	Average Bearer Throughput for uplink (kbps) - based on BDR duration
Total DL (Bytes)	Data volume (in Bytes) received by the MS during this session (IP layer + overhead)
Total UL (Bytes)	Data volume (in Bytes) sent by the MS during this session (IP layer + overhead)
HTTP DL (Bytes)	HTTP data volume (in Bytes) received by the MS during this session
HTTP UL (Bytes)	HTTP data volume (in Bytes) sent by the MS during this session
Social Media DL (Bytes)	Social Media data volume (in Bytes) received by the MS during this session
Social Media UL (Bytes)	Social Media data volume (in Bytes) sent by the MS during this session
YouTube DL (Bytes)	YouTube data volume (in Bytes) received by the MS during this session
YouTube UL (Bytes)	YouTube data volume (in Bytes) sent by the MS during this session
Netflix DL (Bytes)	Netflix data volume (in Bytes) received by the MS during this session
Netflix UL (Bytes)	Netflix data volume (in Bytes) sent by the MS during this session
Google DL (Bytes)	Google data volume (in Bytes) Received by the MS during this session
Google UL (Bytes)	Google data volume (in Bytes) sent by the MS during this session

Field	Description
Email DL (Bytes)	Email data volume (in Bytes) Received by the MS during this session
Email UL (Bytes)	Email data volume (in Bytes) sent by the MS during this session
Gaming DL (Bytes)	Gaming data volume (in Bytes) Received by the MS during this session
Gaming UL (Bytes)	Gaming data volume (in Bytes) sent by the MS during this session
Handset Manufacturer	Handset manufacturer
Subscriber Country	Operator country of the subscriber
Subscriber Group	Operator group of the subscriber
Subscriber Operator	Operator of the subscriber
TCP DL Retrans. Vol (Bytes)	TCP volume of Downlink packets detected as retransmitted (bytes)
TCP DL Vol. (Bytes)	TCP volume transferred in Downlink direction (bytes)
TCP UL Retrans. Vol (Bytes)	TCP volume of Uplink packets detected as retransmitted (bytes)
TCP UL Vol. (Bytes)	TCP volume transferred in Uplink direction (bytes)
UDP DL Vol. (Bytes)	UDP volume transferred in Downlink direction (bytes)
UDP UL Vol. (Bytes)	UDP volume transferred in Uplink direction (bytes)
WAP DL (Bytes)	WAP data volume (in Bytes) received by the MS during this session
WAP UL (Bytes)	WAP data volume (in Bytes) sent by the MS during this session

- sessions frequency
- the duration of the session
- the sessions total traffic (download and upload (bytes))

Engagement metrics

- Average TCP retransmission
- Average RTT
- Handset type
- Average throughput

Experience metrics

- Engagement score
- Experience score

Satisfaction metrics

Non graphical Univariate analysis

- This preliminary data analysis step focuses on four points, i.e These include: measures of central tendency, i.e. the mean, the media and mode, measures of spread, i.e. variability, variants and standard deviation, the shape of the distribution, and the existence of outliers

- Outliers were detected and treated using the Inter Quartile range of each features
- It can be observed that average duration users spent per session is about 83s, the values ranges from 149.748s to as low as 22.33 s
- The average user spends about **474.52 MB** data
- Most of the data spent was on Gaming

Measure	Number of xDR sessions	MSISDN/ Number	Dur. (s)	Total data (MBytes)	Total Google data (MBytes)	Total Social Media data (MBytes)	Total Email data (MBytes)	Total Youtube data (MBytes)	Total Netflix data (MBytes)	Total Gaming data (MBytes)	Total Other data (MBytes)
mean	2.101738	3.37E+10	91.99019	473.4095	7.444549	1.743317	2.152383	21.58196	21.57196	411.0311	409.0946
min	1	3.36E+10	7.146	27.6147	0.038462	0.001491	0.007972	0.075248	0.093872	0.292166	0.488061
25th percentile	1	3.37E+10	54.4055	271.8784	4.712322	0.888987	1.295274	15.25339	15.2496	209.2884	208.5389
50th percentile	1	3.37E+10	86.399	475.1422	7.449817	1.739223	2.155215	21.60171	21.6013	412.7595	409.7008
75th percentile	2	3.37E+10	120.381	674.5478	10.1786	2.602118	3.012089	27.90574	27.90885	612.3263	609.5188
max	11	3.37E+10	244.9	908.5096	14.8094	3.481732	4.308735	43.09662	43.09511	819.3996	819.703

Table 1: Univariate description of the dataset

Graphical Univariate Analysis

- These analysis involves the graphical presentation of the distribution of each features in the dataset
- Figure 1 shows the distribution plot of the Total duration of session, the mean and the percentiles are clustered around a point
- Figure 2 shows the distribution plot of the Total Email data, the distribution has more data greater than the mean

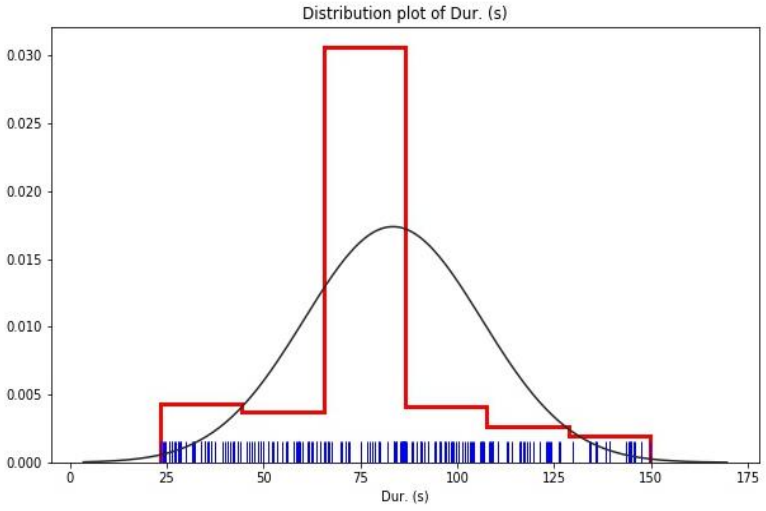


Figure1: Distribution plot of total duration(s)

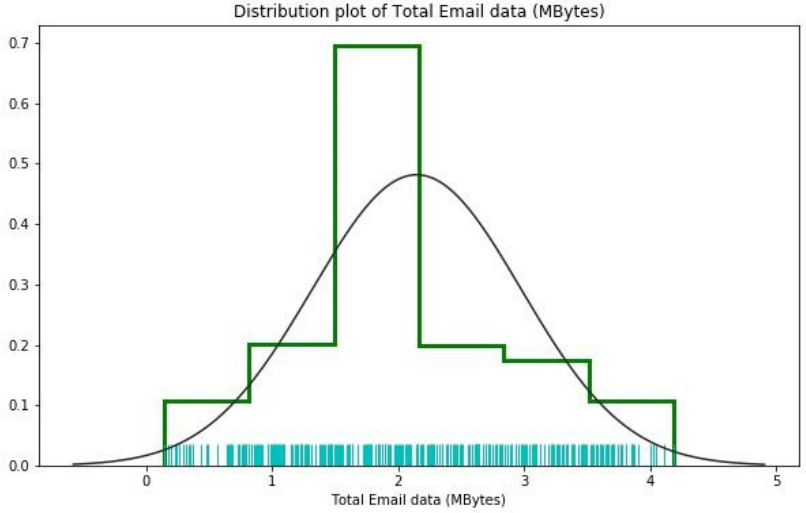


Figure1: Distribution plot of total email data (MB)

Graphical Univariate Analysis cont.

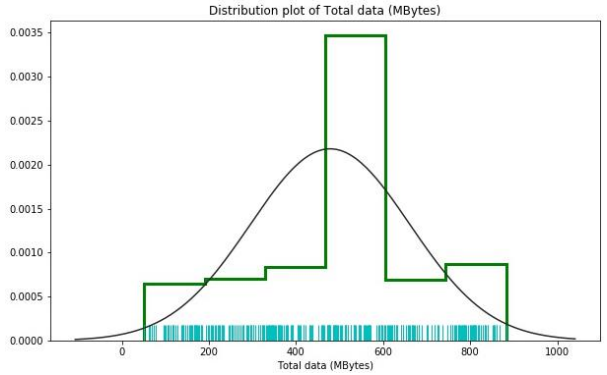


Figure 3: Distribution plot of the Total Data usage

- The plot shows a wide spread amount of data usage over similar ranges

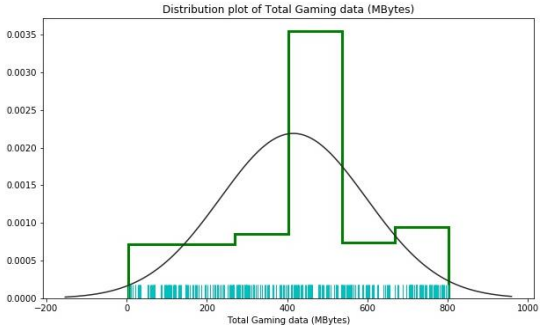


Figure 5:Distribution plot of the Total Gaming Data usage

The plot shows deviation from the normal fitted curve, it shows a spread away from the mean towards left

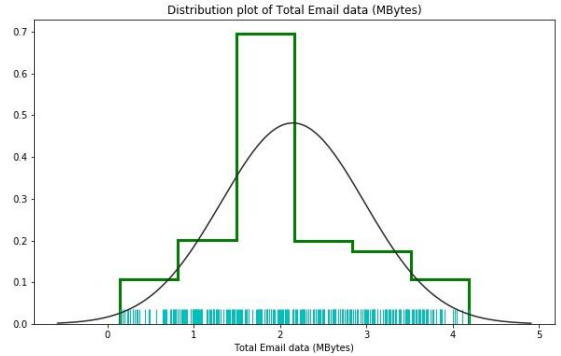


Figure 4: Distribution plot of the Total Email data usage

- The plot is not normally distributed, it has wide spread, with more data greater than the mean

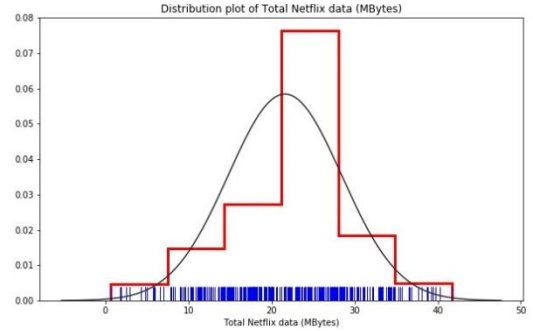


Figure 6:Distribution plot of the Total Netflix Data usage

The plot shows deviation from the normal fitted curve, it shows there are more data less than the mean

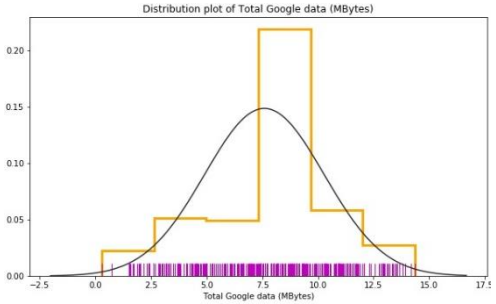


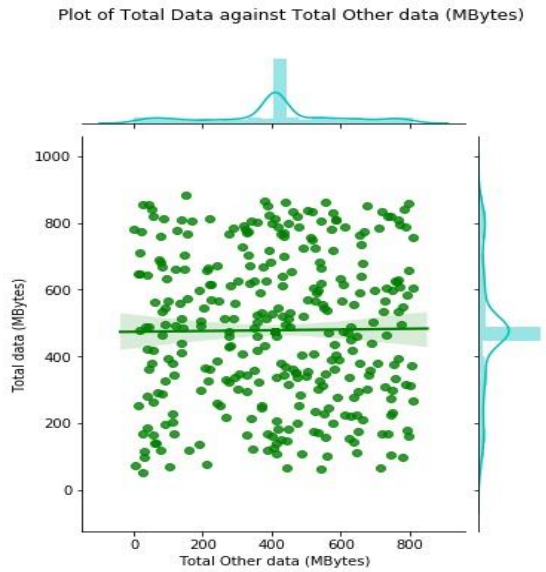
Figure 7:Distribution plot of the Total Google Data usage

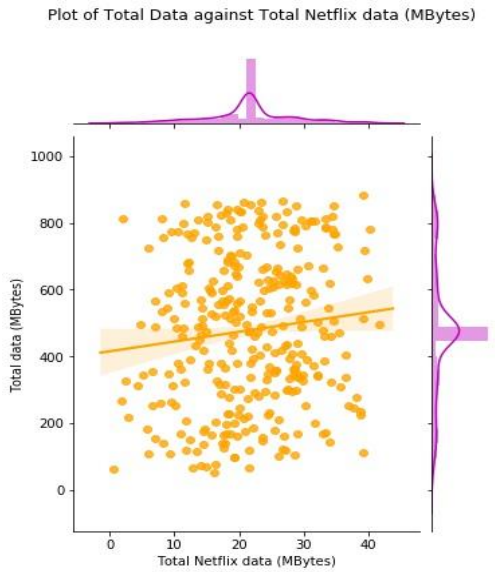
The above plot shows deviation from the normal fitted curve, it shows a spread around a data point that is skewed to the left

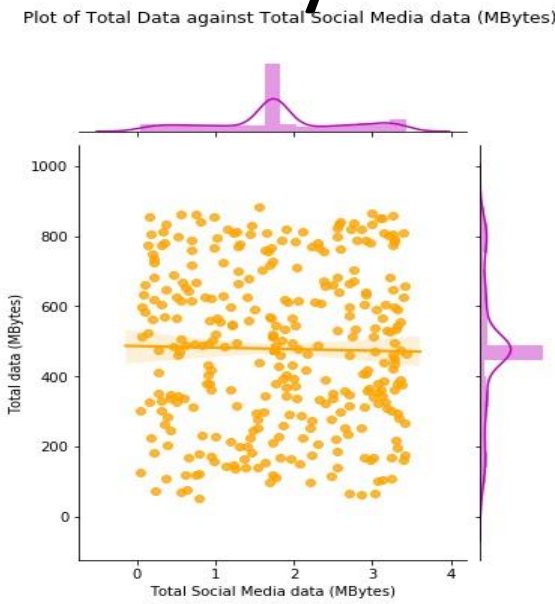
Graphical Bivariate Analysis

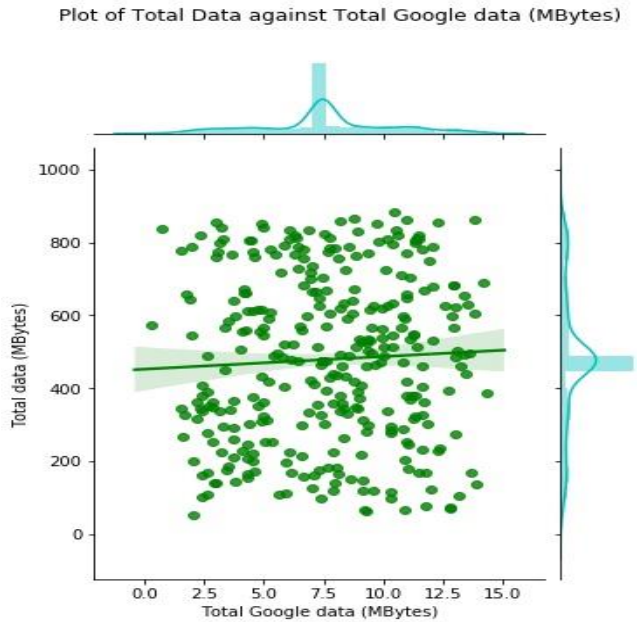
- This examines the relationship between the distribution two features
- In this study, comparison was done between the aforementioned features and the total data
- There is a linear relationship between the total data used and data used during gaming

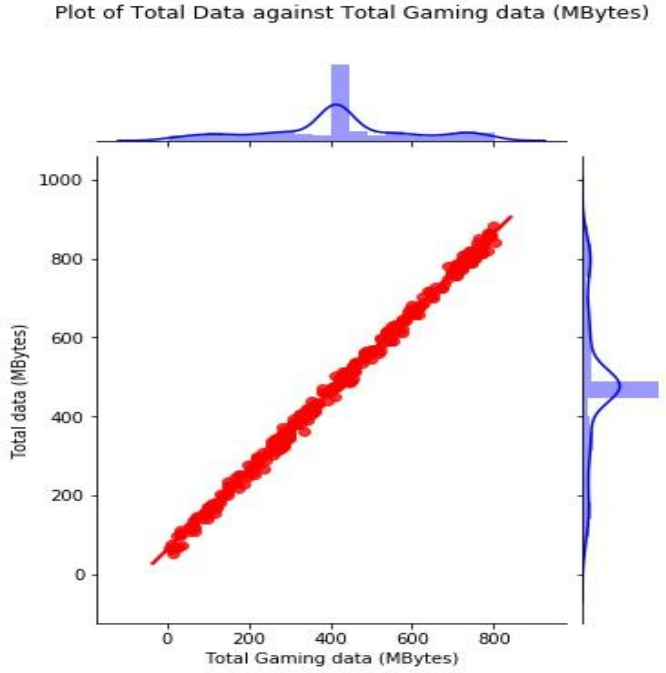
Graphical Bivariate Analysis

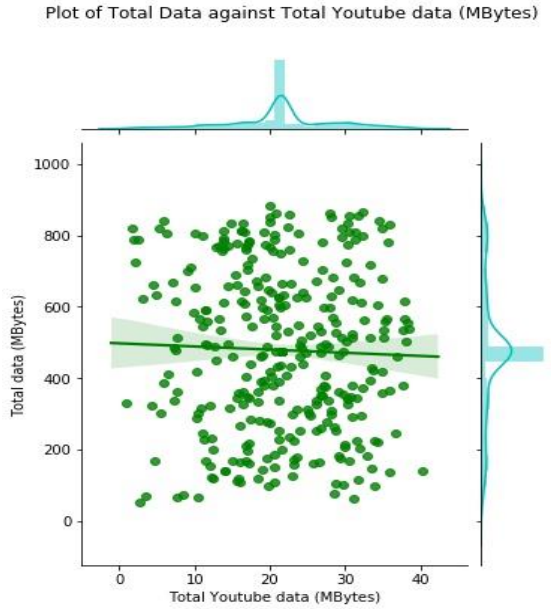


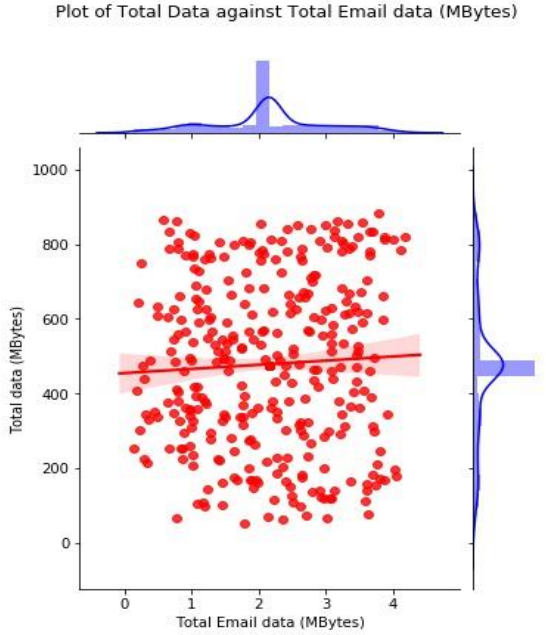












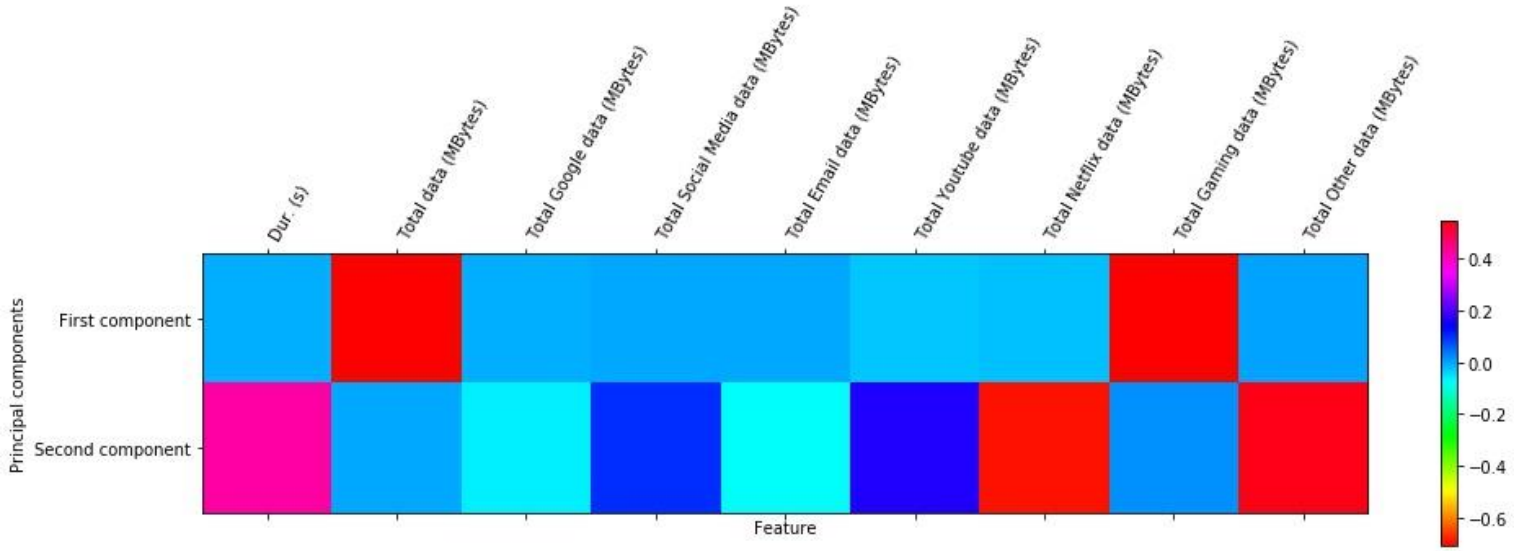
Correlation Analysis

- From the table, it can be observed that there is astrong correlation between total data and total gaming data, while the correlation among other groups are weak (positive and negative)

	Dur. (s)	Total data (MBytes)	Total Google data (MBytes)	Total Social Media data (MBytes)	Total Email data (MBytes)	Total Youtube data (MBytes)	Total Netflix data (MBytes)	Total Gaming data (MBytes)	Total Other data (MBytes)
Dur. (s)	1	0.005798	-0.00216	-0.00753	0.004465	0.004417	-0.00807	0.005923	0.001452
Total data (MBytes)	0.005798	1	0.010404	0.002539	0.005373	0.038423	0.038192	0.998266	-0.00152
Total Google data (MBytes)	-0.00216	0.010404	1	-0.0011	-0.00582	0.004165	0.000829	-0.00406	-0.00324
Total Social Media data (MBytes)	-0.00753	0.002539	-0.0011	1	0.005871	-0.0016	-0.00112	-0.00162	0.004374
Total Email data (MBytes)	0.004465	0.005373	-0.00582	0.005871	1	-0.00047	-0.00014	0.001125	-0.00217
Total Youtube data (MBytes)	0.004417	0.038423	0.004165	-0.0016	-0.00047	1	0.001199	0.000703	0.002411
Total Netflix data (MBytes)	-0.00807	0.038192	0.000829	-0.00112	-0.00014	0.001199	1	0.000449	-0.00854
Total Gaming data (MBytes)	0.005923	0.998266	-0.00406	-0.00162	0.001125	0.000703	0.000449	1	-0.0016
Total Other data (MBytes)	0.001452	-0.00152	-0.00324	0.004374	-0.00217	0.002411	-0.00854	-0.0016	1

Dimensionality Reduction

- The first component has most of the feature to be positive
- The second component was able to retain most of the feature (represente by the colors with high positive)
- The total Netflix data on both components is negative



Top 10 user by engagement metric

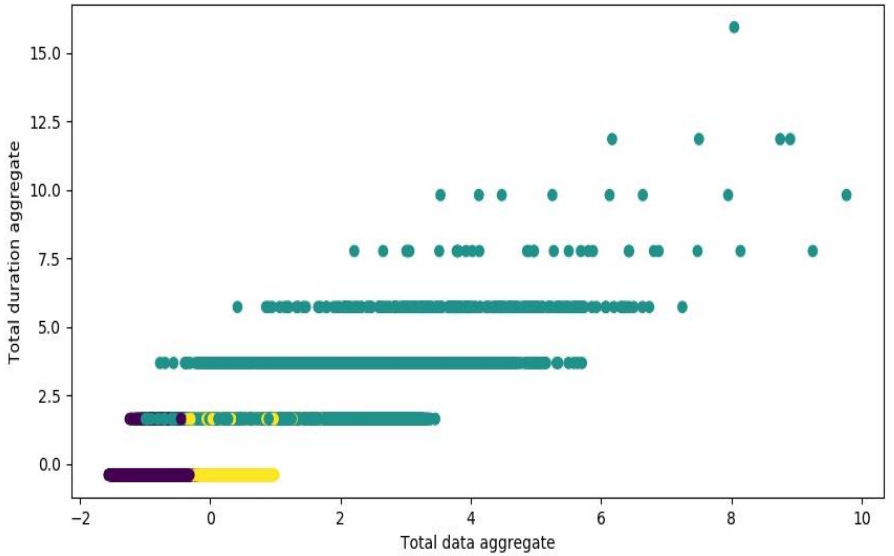
Duration	
MSISDN/Number	Dur. (m)
3.37E+10	12.44339
3.36E+10	12.64114
3.37E+10	12.98228
3.37E+10	13.26967
3.37E+10	13.56621
3.36E+10	13.76465
3.36E+10	14.59329
3.37E+10	15.06969
3.37E+10	16.93144
3.36E+10	17.46782

No of session	
MSISDN/Number	Dur. (m)
3.37E+10	6
3.36E+10	6
3.36E+10	6
3.36E+10	6
3.37E+10	6
3.36E+10	7
3.37E+10	7
3.37E+10	7
3.37E+10	7
3.37E+10	9

Total Data Traffic	
MSISDN/Number	Total Data
3.37E+10	3.2E+09
3.37E+10	3.29E+09
3.37E+10	3.3E+09
3.36E+10	3.46E+09
3.37E+10	3.49E+09
3.36E+10	3.53E+09
3.37E+10	3.75E+09
3.36E+10	3.8E+09
3.37E+10	3.93E+09
3.37E+10	4.11E+09

Cluster analysis for engagement metrics

- The third cluster had the highest number of data point
- However, the first cluster contain user with high amount of data usage



First cluster			
	Total data	Duration	count
count	22413	22413	22413
mean	1.03E+17	1.808135	1.010931
std	4.43E+16	0.90082	0.103981
min	1.20E+16	0.121553	1
25%	6.49E+16	1.190434	1
50%	1.03E+17	1.708496	1
75%	1.41E+17	2.490358	1
max	1.88E+17	4.081366	2

Second Cluster			
	Total data	Duration	count
count	7925	7925	7925
mean	1.13E+09	4.126434	2.21489
std	4.27E+08	1.700625	0.523434
min	1.6E+08	0.391158	2
25%	8.44E+08	2.983302	2
50%	1.1E+09	4.05078	2
75%	1.38E+09	5.084551	2
max	4.11E+09	17.46782	9

Third cluster			
	Total data	Duration	count
count	22921	22921	22921
mean	7.12E+08	1.739552	1.032808
std	1.27E+08	0.894512	0.178139
min	4.62E+08	0.123772	1
25%	6.05E+08	1.049135	1
50%	7.1E+08	1.651249	1
75%	8.16E+08	2.394022	1
max	1.34E+09	4.081667	2

TOP 10 USER PER APPLICATION

EMAIL	
MSISDN/ Number	Total Email data (Bytes)
3.37E+10	18857830
3.37E+10	18951528
3.37E+10	19050869
3.37E+10	19235744
3.37E+10	19489977
3.37E+10	19578644
3.37E+10	20610254
3.37E+10	20999792
3.37E+10	21923171
3.37E+10	22923413

GAMING	
MSISDN/ Number	Total Gaming data (Bytes)
3.37E+10	3.71E+09
3.37E+10	3.74E+09
3.37E+10	3.78E+09
3.37E+10	3.96E+09
3.37E+10	3.97E+09
3.37E+10	4.11E+09
3.36E+10	4.12E+09
3.37E+10	4.19E+09
3.37E+10	4.25E+09
3.37E+10	4.55E+09

GOOGLE	
MSISDN/ Number	Total Google data (Bytes)
3.37E+10	65332551
3.37E+10	65437450
3.37E+10	67375921
3.37E+10	69340837
3.37E+10	69585701
3.37E+10	71974850
3.37E+10	77824684
3.37E+10	78719572
3.36E+10	85492317
3.37E+10	87634510

NETFLIX	
MSISDN/ Number	Total Netflix data (Bytes)
3.37E+10	1.91E+08
3.37E+10	1.93E+08
3.36E+10	1.93E+08
3.37E+10	1.99E+08
3.36E+10	2.01E+08
3.37E+10	2.01E+08
3.37E+10	2.05E+08
3.37E+10	2.05E+08
3.37E+10	2.12E+08
3.37E+10	2.34E+08

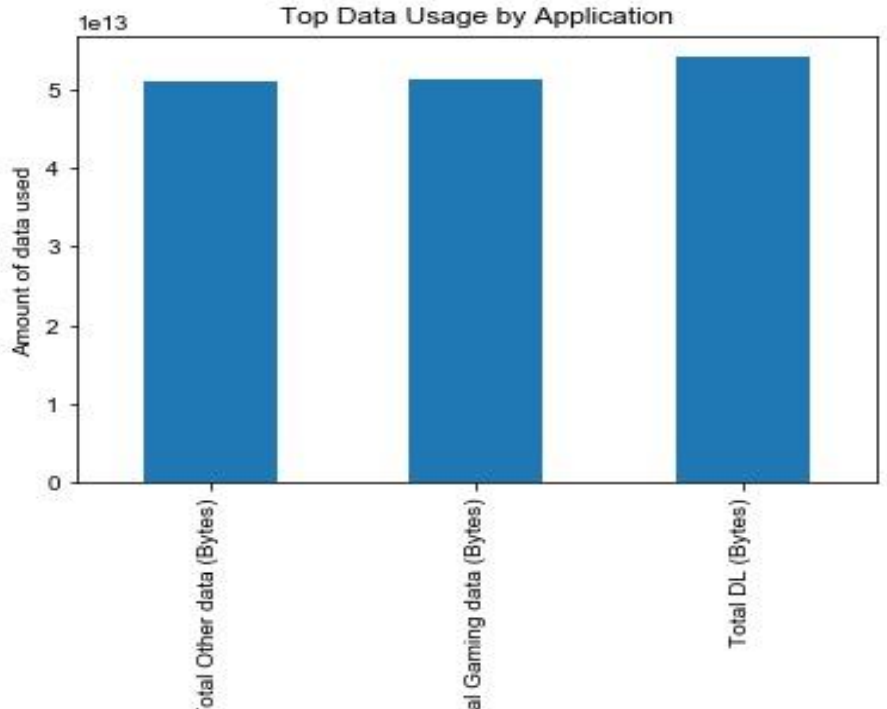
OTHER DATA	
MSISDN/ Number	Total Other data (Bytes)
3.37E+10	3.64E+09
3.37E+10	3.7E+09
3.37E+10	3.72E+09
3.37E+10	3.94E+09
3.37E+10	4.09E+09
3.37E+10	4.14E+09
3.37E+10	4.21E+09
3.36E+10	4.25E+09
3.37E+10	4.55E+09
3.37E+10	4.61E+09

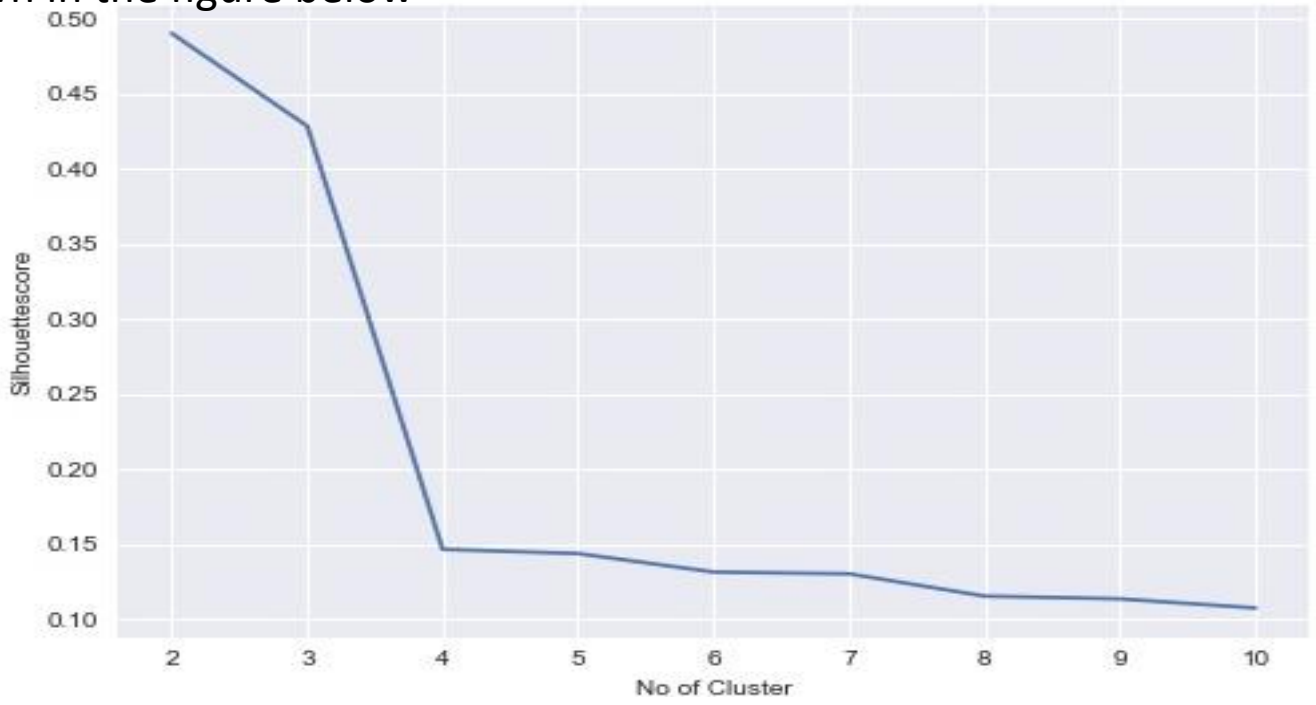
SOCIAL MEDIA	
MSISDN/ Number	Total Social Media data (Bytes)
3.37E+10	15967607
3.37E+10	16105635
3.37E+10	16179085
3.36E+10	16464155
3.37E+10	16585183
3.37E+10	17311846
3.37E+10	17982623
3.37E+10	18055133
3.37E+10	19069454
3.37E+10	23800834

YOUTUBE	
MSISDN/ Number	Total Youtube data (Bytes)
3.37E+10	1.88E+08
3.37E+10	1.89E+08
3.37E+10	1.91E+08
3.37E+10	1.93E+08
3.37E+10	1.96E+08
3.36E+10	2.01E+08
3.36E+10	2.29E+08
3.37E+10	2.36E+08
3.37E+10	2.37E+08
3.37E+10	2.49E+08

Analysis of Engagement metrics

- The optimum number of cluster was obtained from the silhouette score as shown in the figure below
- The optimum number of cluster is 2
- The top usage per application is also shown in the figure below





10 of the top, bottom and most frequent experience metrics

Most frequent Avg Bearer	
MSISDN/Number	Frequency
3.37E+10	6
3.36E+10	6
3.36E+10	6
3.36E+10	6
3.37E+10	6
3.36E+10	7
3.37E+10	7
3.37E+10	7
3.37E+10	7
3.37E+10	9

Top Total Avg Bearer	
MSISDN/Number	sum
3.36E+10	8871
3.36E+10	15
3.36E+10	67
3.36E+10	126
3.36E+10	489
3.36E+10	113
3.36E+10	15
3.36E+10	483
3.36E+10	602
3.36E+10	133

Bottom Total Avg Bearer	
MSISDN/Number	sum
3.37E+10	126
3.37E+10	96
3.37E+10	84
3.37E+10	172
3.37E+10	7185
3.37E+10	15
3.37E+10	67
3.37E+10	15
3.37E+10	8192
3.37E+10	15

Most Frequent Total Avg RTT	
MSISDN/Number	Frequency
3.37E+10	6
3.36E+10	6
3.36E+10	6
3.36E+10	6
3.37E+10	6
3.36E+10	7
3.37E+10	7
3.37E+10	7
3.37E+10	7
3.37E+10	9

Bottom Total Avg RTT	
MSISDN/Number	sum
3.37E+10	44.44266
3.37E+10	61
3.37E+10	90
3.37E+10	76
3.37E+10	76.44266
3.37E+10	44.44266
3.37E+10	39
3.37E+10	44.44266
3.37E+10	29
3.37E+10	44.44266

Top Total Avg RTT	
MSISDN/Number	sum
3.36E+10	75.44266
3.36E+10	44.44266
3.36E+10	27
3.36E+10	94
3.36E+10	127
3.36E+10	28
3.36E+10	44.44266
3.36E+10	77
3.36E+10	124
3.36E+10	88.88532

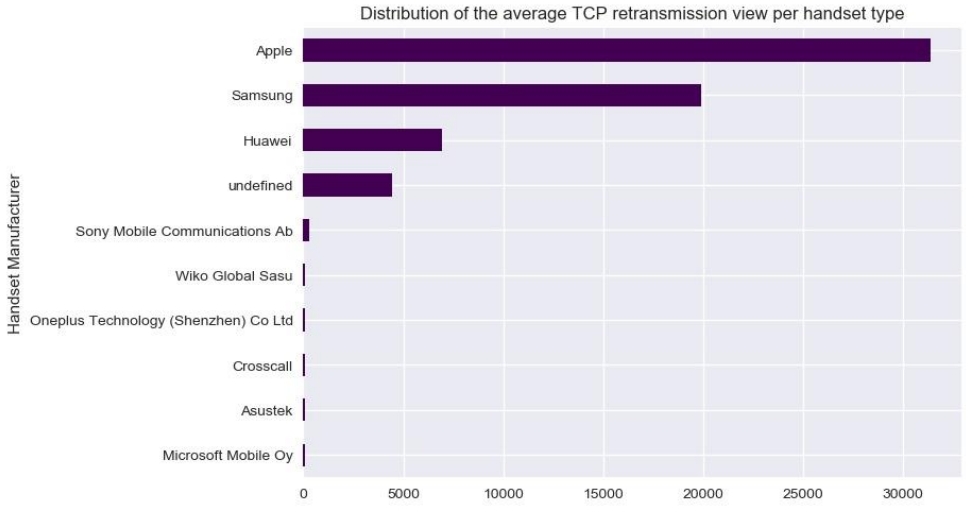
Most frequent Total TCP DL Retrans. Vol (Bytes)	
MSISDN/Number	Frequency
3.37E+10	6
3.36E+10	6
3.36E+10	6
3.36E+10	6
3.37E+10	6
3.36E+10	7
3.37E+10	7
3.37E+10	7
3.37E+10	7
3.37E+10	9

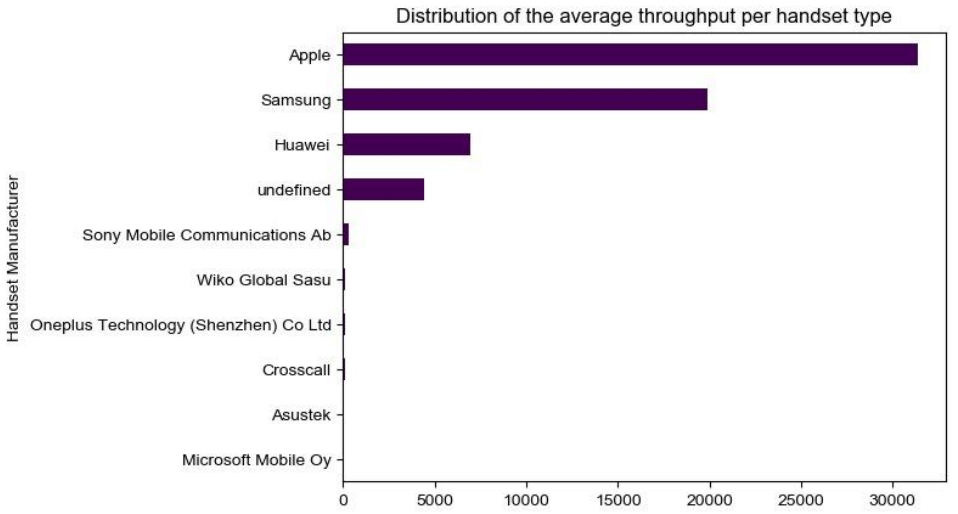
Top Total TCP DL Retrans. Vol (Bytes)	
MSISDN/Number	sum
3.36E+10	574330.6
3.36E+10	287165.3
3.36E+10	287165.3
3.36E+10	574330.6
3.36E+10	4087
3.36E+10	287165.3
3.36E+10	287165.3
3.36E+10	287165.3
3.36E+10	287165.3
3.36E+10	574330.6

Bottom Total Total TCP DL Retrans. Vol (Bytes)	
MSISDN/Number	sum
3.37E+10	287165.3
3.37E+10	287165.3
3.37E+10	287165.3
3.37E+10	574330.6
3.37E+10	574330.6
3.37E+10	287165.3
3.37E+10	287165.3
3.37E+10	287165.3
3.37E+10	287165.3
3.37E+10	287165.3

Distribution of the average throughput and average TCP retransmission view per handset type

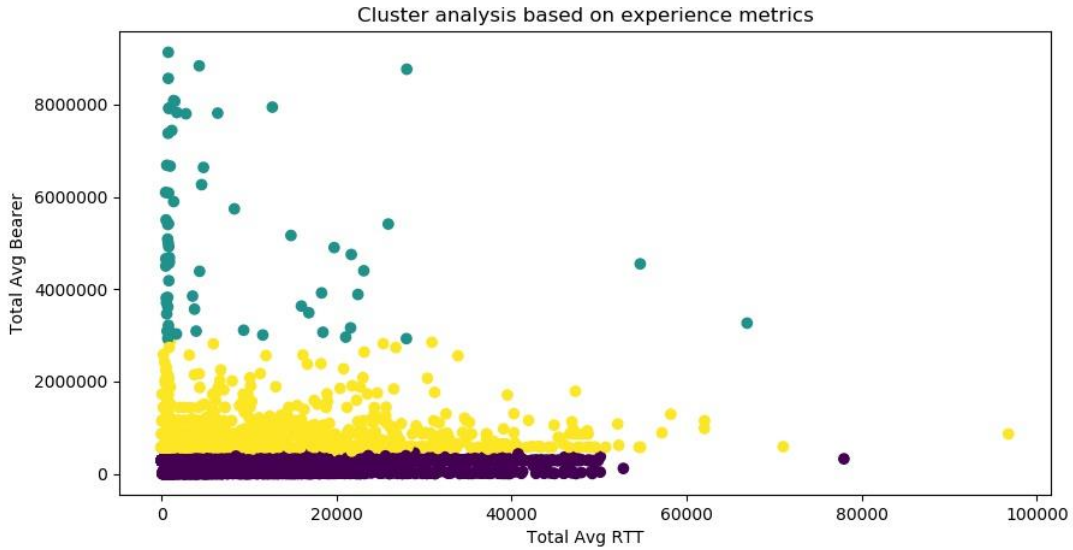
- Apple handset have the most usage as shown by the two charts





Cluster analysis of the experience metrics

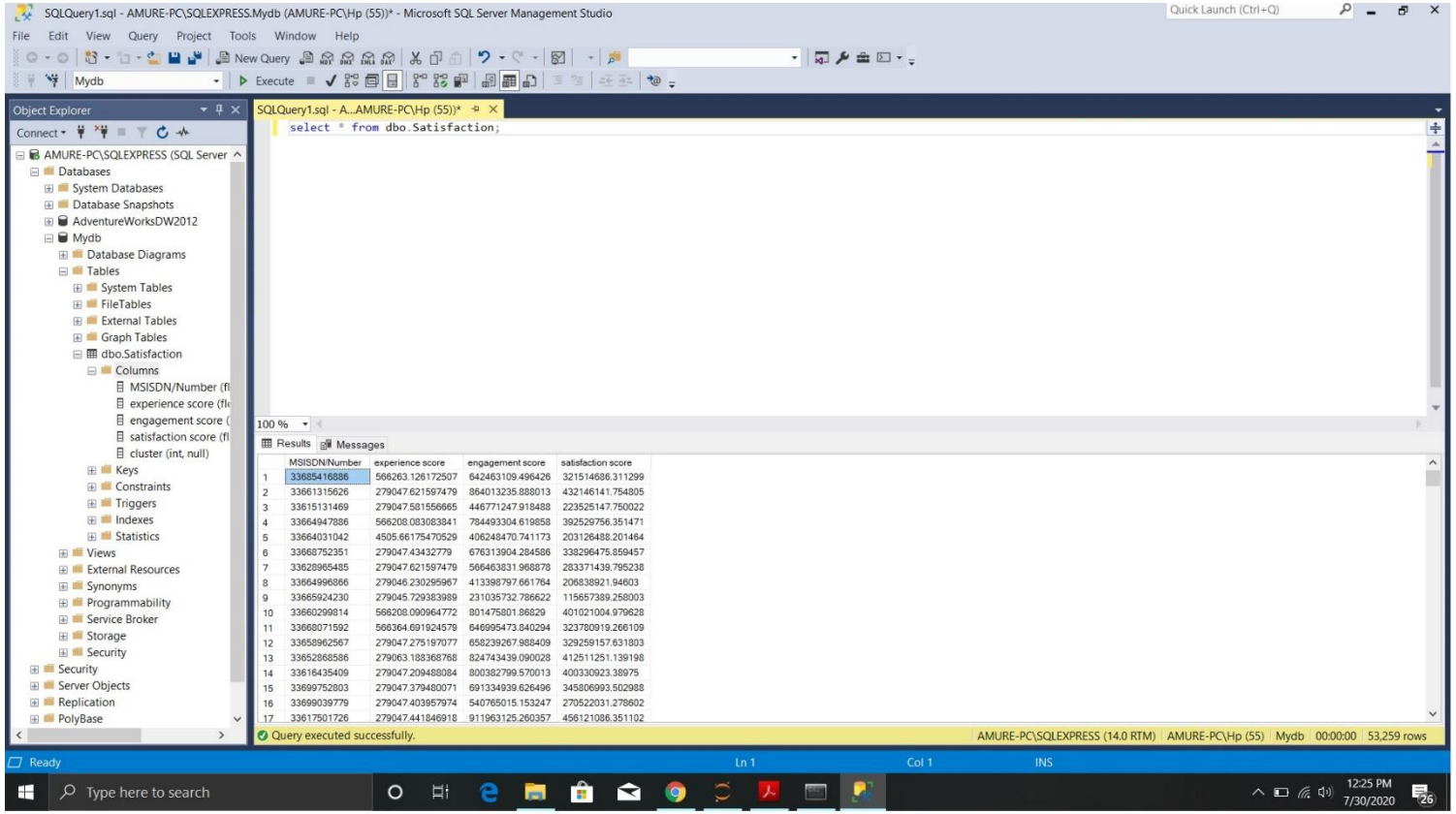
- Cluster 1 had about 8537 data point, it has the highest Total Avg RTT
- The second cluster has a relatively lower value across the features
- The third cluster has the highest Total TCP DL Retrans. Vol (Bytes)



Satisfaction Metrics

MSISDN/Number	Satisfaction score
3.36E+10	1.6E+09
3.37E+10	1.65E+09
3.36E+10	1.65E+09
3.37E+10	1.73E+09
3.37E+10	1.75E+09
3.37E+10	1.76E+09
3.37E+10	1.87E+09
3.37E+10	1.9E+09
3.37E+10	1.96E+09
3.37E+10	2.06E+09

DATABASE


SQLQuery1.sql - AMURE-PC\SQLEXPRESS.Mydb (AMURE-PC\Hp (55))* - Microsoft SQL Server Management Studio
File Edit View Query Project Tools Window Help
New Query
Mydb
Execute
Object Explorer
Connect
AMURE-PC\SQLEXPRESS (SQL Server
Databases
System Databases
Database Snapshots
AdventureWorksDW2012
Mydb
Database Diagrams
Tables
System Tables
FileTables
External Tables
Graph Tables
dbo.Satisfaction
Columns
MSISDN/Number (fl
experience score (fl
engagement score (
satisfaction score (fl
cluster (int, null)
Keys
Constraints
Triggers
Indexes
Statistics
Views
External Resources
Synonyms
Programmability
Service Broker
Storage
Security
Security
Server Objects
Replication
PolyBase
SQLQuery1.sql - A...AMURE-PC\Hp (55))*
select * from dbo.Satisfaction;
100 %
Results
Messages
MSISDN/Number experience score engagement score satisfaction score
1 33685416886 566263.126172507 642463109.496426 321514686.311299
2 33661315626 279047.621597479 864013235.888013 432146141.754805
3 33615131469 279047.581556665 446771247.918488 223525147.750022
4 33664947886 566208.083083841 784493304.619858 392529756.351471
5 33664031042 4505.66175470529 406248470.741173 203126488.201464
6 33668752351 279047.43432779 676313904.284586 338296475.859457
7 33628965485 279047.621597479 566463831.968878 283371439.795238
8 33664996866 279046.230295967 413398797.661764 206838921.94603
9 33665924230 279045.729383989 231035732.786622 115657389.258003
10 33660299814 566208.090964772 801475801.86829 401021004.979628
11 33668071592 566364.691924579 646995473.840294 323780919.266109
12 33658962567 279047.275197077 658239267.988409 329259157.631803
13 33652868586 279063.188368768 824743439.090028 412511251.139198
14 33616435409 279047.209488084 800382799.570013 400330923.38975
15 33699752803 279047.379480071 691334939.626496 345806993.502988
16 33699039779 279047.403957974 540765015.153247 270522031.278602
17 33617501726 279047.441846918 911963125.260357 456121086.351102
Query executed successfully.
AMURE-PC\SQLEXPRESS (14.0 RTM) AMURE-PC\Hp (55) Mydb 00:00:00 53,259 rows
Ready
Ln 1 Col 1 INS
Type here to search
12:25 PM
7/30/2020

Conclusion and Recommendation

- There is a positive correlation between the amount data used in gaming and data by the users, this means majority of the data is consumed by this group, although there are more people in this category whose usage is below average
- Quite a number of people tend to use their data for Email services
- An average user has about 2 sessions and not likely to go above 11 sessions
- The users can be sufficiently grouped with the experience and engagement metrics
- From this grouping, large groups representing highly engaged users that are satisfied are observed
- From all the analysis carried out, it can be concluded that the company has viable prospects