USER ANALYTICS IN THE TELECOMMUNICATION INDUSTRY (TellCo).

Overview:
Analysing users’ experience, engagement and satisfaction to check for growth opportunities in TellCo and making recommendation on whether it is worth buying or selling.

User overview analysis

Relevant Features:

- The data provided has numerous features but not all are worth exploring. The following are some of the relevant features:
  - ***Bearer Id***: The xDR session identifier. It's a categorical feature.
  - ***Dur. (ms).1****:* This represents the session duration in microseconds. It's a numerical feature.
  - ***MSISDN/Number****:* This is the unique customer identifier(number). It's numerical but can be treated as a categorical feature due to it's uniqueness to each customer.
  - ***Social Media DL (Bytes)*** and ***Social Media UL (Bytes)****:* Social media data volume in bytes received and sent respectively during the session. Its a numerical feature of type float.

    NB: ** The description above extends to the following features: *YouTube DL (Bytes), YouTube UL (Bytes), Netflix DL (Bytes), Netflix UL (Bytes), Google DL (Bytes), Google UL (Bytes), Email DL (Bytes), Email UL (Bytes), Gaming DL (Bytes), Gaming UL (Bytes), Other DL, Other UL.*

    The only difference is in the application type.

User overview analysis

- ***Avg RTT DL (ms)*** and ***Avg RTT UL (ms):*** Average Round Trip Time measurement Downlink and Uplink direction in microseconds. Both are numerical features of type float.
- ***Avg Bearer TP DL (kbps)*** and ***Avg Bearer TP UL (kbps):*** Average Bearer Throughput for Downlink and uplink in kbps based on BDR duration. Both are numerical features of type float.
- ***TCP DL Retrans. Vol (Bytes)*** and ***TCP DL Retrans. Vol (Bytes)***: TCP volume of Downlink and Uplink packets detected as retransmitted in bytes. Both are numerical features of type float.
- ***Handset Type***: The handset type of the mobile device.It's a categorical feature of type object.
- ***Handset Manufacturer:*** The handset manufacturer. It's a categorical feature of type object.
- ***Total DL (Bytes)*** and ***Total UL (Bytes):*** Total data volume in bytes received or sent during the session. It's a numerical feature of type float.

User overview analysis

The following metrics wil be used for the analysis:

- **User engagement :**

  This is calculated as the shortest distance between an observation and a cluster centroid of the least engaged cluster.

  The clusters will be generated from performing a KMeans cluster analysis on the *sessions duration, sessions traffic* and *sessions frequency* aggregated per customer.

  This metric will give a picture of how intense a customer's engagement with various services is and thus provide directions on what services and products to put more effort towards in order to increase the engagement further and encourage more customers.

- **User experience:**

  The metric is calculated the same way as user engagement, only that the KMeans analysis is performed on the *average TCP transmission, average RTT, average throughput* and *handset-type count* aggregated per user.

  This will help in the optimization of products and services so that it meets the evolving users needs and expectations.

- **User satisfaction:**

  Satisfaction is calculated as the average of a user's engagement and experience scores.

  This will give an overall view of a user's satisfaction in all aspects. Most satisfied users' behaviours can be observed and the services and products contributing to the set of behaviours be given a higher priority and the insignificant ones be rid off or improved.

User overview analysis

Non-graphical univariate analysis.

Figure 1.0

	count	mean	std	min	25%	50%	75%	max
Dur. (ms).1	150001.0	1.046091e+08	8.103734e+07	7142988.0	57442058.0	8.639998e+07	1.324307e+08	1.859336e+09
Total UL (Bytes)	150001.0	4.112121e+07	1.127635e+07	2866892.0	33222029.0	4.114324e+07	4.903424e+07	7.833131e+07
Total DL (Bytes)	150001.0	4.546434e+08	2.441421e+08	7114041.0	243107173.0	4.558409e+08	6.657051e+08	9.029696e+08
social_media	150001.0	1.828250e+06	1.035646e+06	1563.0	932218.0	1.826471e+06	2.727487e+06	3.650861e+06
netflix	150001.0	2.262861e+07	9.260820e+06	98432.0	15979455.0	2.263554e+07	2.929044e+07	4.519815e+07
gaming	150001.0	4.303331e+08	2.440199e+08	306358.0	218727939.0	4.316150e+08	6.414159e+08	8.592028e+08
youtube	150001.0	2.264348e+07	9.246800e+06	78903.0	15998463.0	2.266177e+07	2.929260e+07	4.519008e+07
google	150001.0	7.807295e+06	3.516420e+06	40330.0	4943599.0	7.812835e+06	1.068280e+07	1.552878e+07
email	150001.0	2.259102e+06	1.071109e+06	8359.0	1359344.0	2.263567e+06	3.159818e+06	4.518036e+06
other	150001.0	4.293653e+08	2.432681e+08	149045.0	218553417.0	4.299865e+08	6.399275e+08	8.595209e+08
Avg RTT DL (ms)	150001.0	1.097957e+02	5.593426e+02	0.0	35.0	5.400000e+01	1.097957e+02	9.692300e+04
Avg RTT UL (ms)	150001.0	1.766288e+01	7.652993e+01	0.0	3.0	7.000000e+00	1.766288e+01	7.120000e+03
Avg Bearer TP DL (kbps)	150001.0	1.330005e+04	2.397180e+04	0.0	43.0	6.300000e+01	1.971000e+04	3.781600e+05
Avg Bearer TP UL (kbps)	150001.0	1.770429e+03	4.625340e+03	0.0	47.0	6.300000e+01	1.120000e+03	5.861300e+04
TCP DL Retrans. Vol (Bytes)	150001.0	2.080991e+07	1.172356e+08	2.0	1332932.0	2.080991e+07	2.080991e+07	4.294426e+09
TCP UL Retrans. Vol (Bytes)	150001.0	7.596587e+05	1.577616e+07	1.0	63009.0	7.596587e+05	7.596587e+05	2.908226e+09

User overview analysis

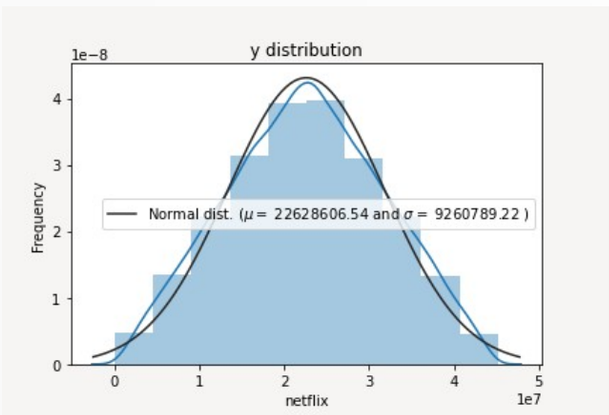
- The **figure 1.0** in the previous slide summarises the statistics of the quantitative variables.
  - The **min** value is the smallest value while the **max** is the largest value in the variable's observations.
  - **50%** is the middle element (median) when the data is arranged in order of magnitude while **25%** and **75%** are the median values for the top and bottom halves of the data.
  - **mean** value is the arithmetic mean which is the central value of a discrete set of numbers.
  - **Std** is the dispersion of a dataset relative to its mean value.

- The measures displayed vary across the features. This can be explained by the different units of measurement across some of them. For the variables where the unit of measurement is constant, for instance, the applications, the variation can be explained by different levels of usage by the users.

User overview analysis

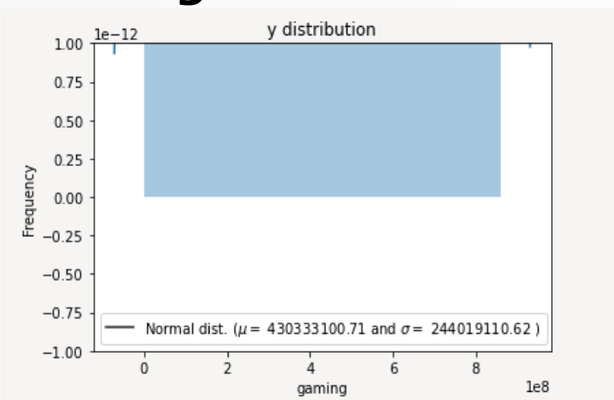
Graphical univariate analysis.

Netflix



The plot is somewhat symmetric and bell-shaped indicating Normally distributed unimodal data with a mean of 22628606 which matches the non-graphical analysis.

Gaming

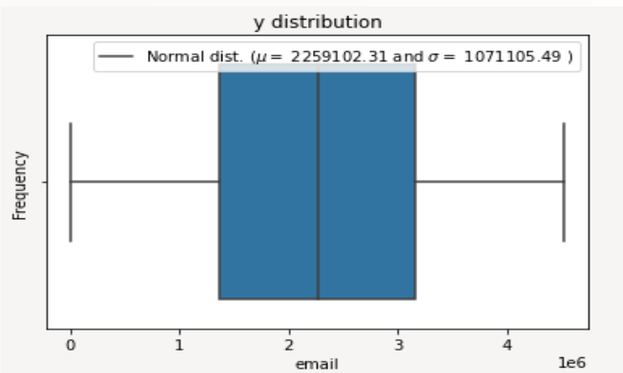


The plot is uniformly shaped indicating consistent multimodal uniformly distributed data with a mean of 430333100 and a standard deviation of 244019110 matching the non-graphical analysis.

User overview analysis

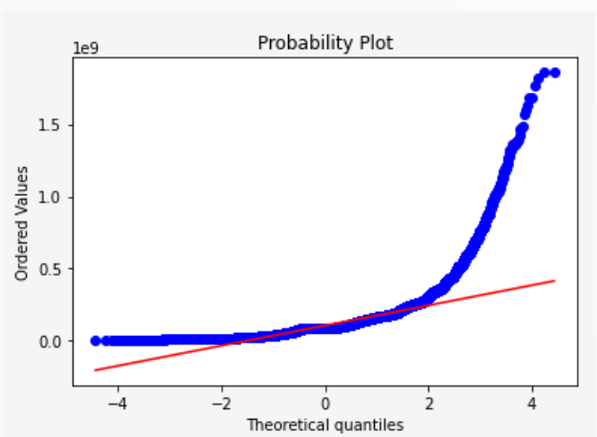
Graphical univariate analysis.

Email



The box plot is symmetrical (median splitting the plot equally) indicating unimodal normally distributed data with a mean of 2259102 and standard deviation of 1071105 bytes.

Session Duration:

- 

A qqplot indicating heavily skewed right-tailed normal distribution.

User overview analysis

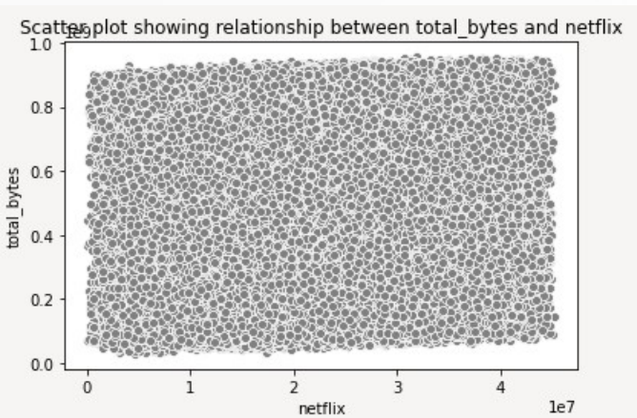
- The plots in the previous slides show the unique distribution shapes obtained from the features.
- Summary of the distributions of the remaining features :
  - **Google**: Unimodal normally distributed data.
  - **Social media**: Multimodal uniformly distributed data.
  - **Total bytes**: Multimodal uniformly distributed data.
  - **Total RTT:** Right skewed normally distributed data.
  - **Total TCP**: Right skewed normally distributed data.
  - **Youtube:** A symmetrical bell-shaped plot indicating normally distributed data.
  - **Total Throughput:** A plot indicating right-tailed unimodal normally distributed data.
- A **Normal** distribution indicates that in a particular feature, majority of the observations are near the mean.
- **Uniform** distribution indicates equal outcome of all observations in a particular feature.

User overview analysis

Bivariate analysis between the apps and the total bytes.

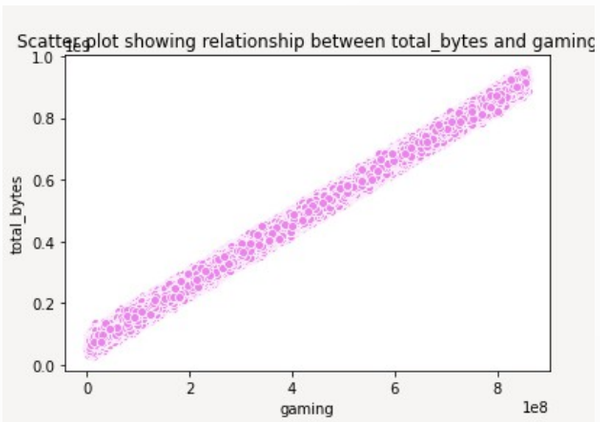
- *Correlation* is a measure of the strength of a linear relationship between two quantitative variable

Netflix



No correlation between netflix and total bytes.

Gaming

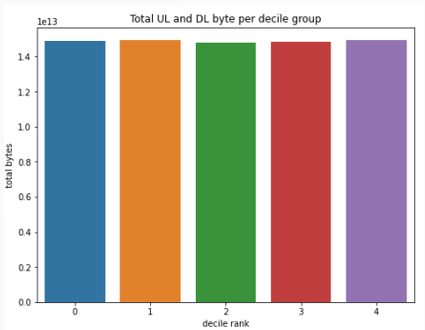


Positive correlation between gaming and total bytes. Could be as a result of high volume of gaming data compared to other apps.

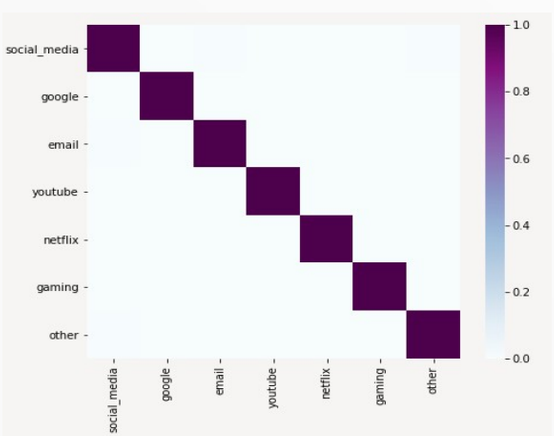
- For the remaining applications: *Email, youtube, google, other and social media*, the plots resemble the netflix plot indicating no correlation between the appplications and the *total bytes.*

User overview analysis

- The plot below shows total data per **decile** class. The classes are generated based on the total sessions duration.



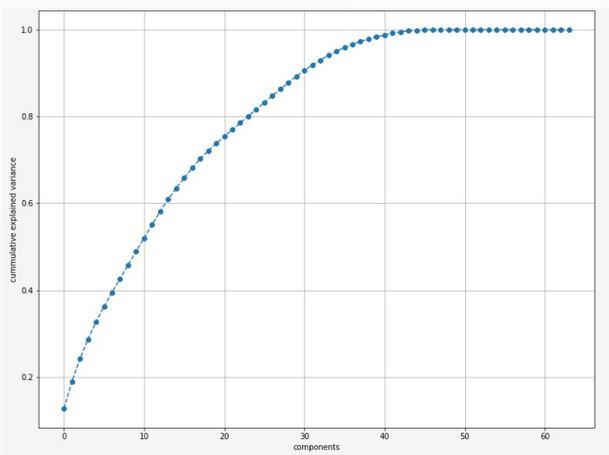
- The total bytes is around the same across the decile groups, this indcates that:
  - The amount of total bytes is independent of session duration.
  - If explanation one is incorrect, then the variation of session duration across the decile groups is minimal.
- The heatmap below shows that the applications have no **correlation** (linear relationship) with each other.



User overview analysis

Dimensionality reduction.

- A cummulative explained variance plot indicating optimal number of components(features) to be 23 after considering the 80% cummulative variance coverage rule.
- This is a massive reduction from the original features.



10 users with the highest number of sessions, highest sessions traffic and longest sessions duration:

- After identifying top 10 users in each of the 3 categories, the following appear in all the 3:
  - *33625779332.0, 33626320676.0, 33760413819.0, 33614892860.0, 33760536639.0*
- The products and services used by the 5 users that appear in the 3 categories can be identified and their special features extended to other products and services in order to trap more users attention.

User engagement analysis

A summary of clusters generated from performing a KMeans analysis on the users based on the sessions count, sessions duration and sessions traffic.

- Cluster 1

	sessions_freq	sessions_duration(ms)	sessions_traffic(bytes)
count	22466.000000	2.246600e+04	2.246600e+04
mean	1.109855	3.157811e+07	5.657125e+08
std	0.314985	1.329984e+07	2.959322e+08
min	1.000000	7.142988e+06	3.802236e+07
25%	1.000000	2.082711e+07	3.308632e+08
50%	1.000000	3.032519e+07	5.594369e+08
75%	1.000000	4.139189e+07	7.677527e+08
max	4.000000	6.426047e+07	1.895711e+09

Cluster 2

	sessions_freq	sessions_duration(ms)	sessions_traffic(bytes)
count	29494.000000	2.949400e+04	2.949400e+04
mean	2.256289	2.887818e+08	1.223616e+09
std	6.394298	5.567422e+08	3.187913e+09
min	1.000000	5.447539e+07	1.770069e+08
25%	2.000000	1.727915e+08	8.213283e+08
50%	2.000000	2.257546e+08	1.059776e+09
75%	3.000000	3.315747e+08	1.444668e+09
max	1084.000000	8.134348e+10	5.397159e+11

Cluster 3

	sessions_freq	sessions_duration(ms)	sessions_traffic(bytes)
count	54896.000000	5.489600e+04	5.489600e+04
mean	1.066016	1.177628e+08	4.657273e+08
std	0.249629	5.129752e+07	2.360029e+08
min	1.000000	4.285693e+07	3.324901e+07
25%	1.000000	8.639992e+07	2.666905e+08
50%	1.000000	9.823675e+07	4.547752e+08
75%	1.000000	1.413285e+08	6.527742e+08
max	3.000000	1.035262e+09	1.214536e+09

- **Session frequency:** The minimum values are the same across the clusters since 1 is the least possible value of session frequency. The maximum values vary with the highest session frequency falling under cluster 2. The average values also vary depending on the the values of the sessions-frequency under each cluster
- **Sessions Duration:** The minimum values have a slight variation across the clusters. The maximum values' variation is a bit wider, with the highest value falling under cluster 2. This could be due to the presence of the highest session-frequency in the cluster assuming correlation between sessions frequency and duration. The same explanation extends to the mean values.
- **Sessions Traffic**: The behaviour observed in the sessions duration is replicated in sessions traffic too. The explanation stands: the values in cluster 2 are overally higher than those in 1 and 3 due to the presence of the highest session-frequency value in the cluster assuming a correlation between sessions frequency and sessions traffic.

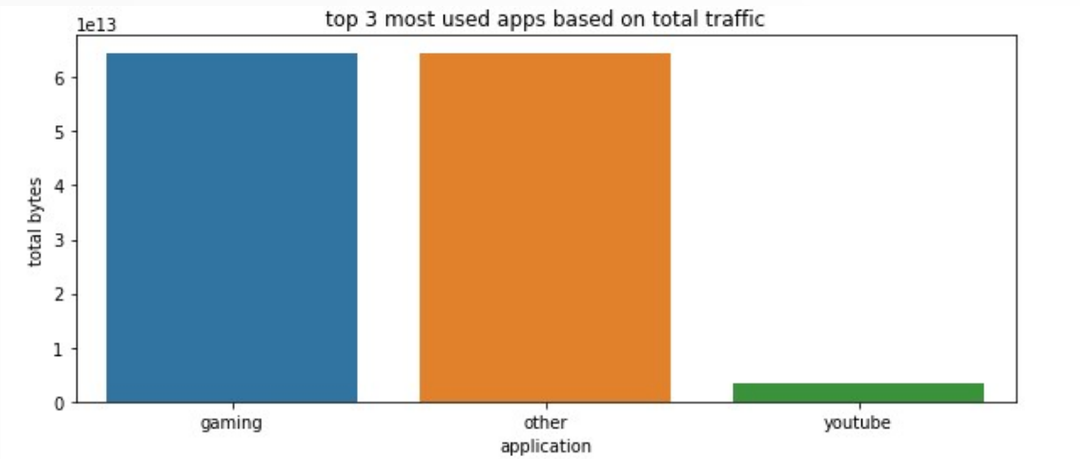
User engagement analysis

Top 10 most engaged users per application:

- After identfying top 10 most enaged users in the applications, the following appear in all 7 of them: *33625779332.0, 33614892860.0, 33626320676.0.*
- To replicate intense engagement in other customers like the 3, the following can be considered:
  - Extending the unique features of the Handset types used by the 3 to other handsets.
  - Increasing the stock of the handsets used by the 3 to force customers into buying them.
- The above recommendations assumes the highest contributor towards a user's engagement with all the applications is the handset type.
- There are other factors that contribute to this, e.g User's interests, most which are tailored to each user, but based on the data provided, Handset Type is the most reasonable contributor.

User engagement analysis

Top 3 most used applications based on sessions total traffic:

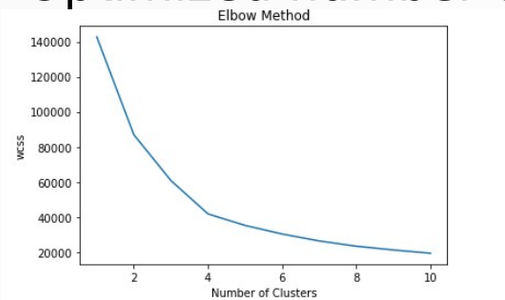


- The large percentage of the '**other**' application can be explained by the many minor applications being represented by the one column.
- As for '**gaming'**, it could be due to the following reasons:
  - Most of the gaming applications consume a lot of bytes compared to other applications.
  - The games have a bigger percentage in the total applications available in a handset.
  - Most of the handset users are gamers.

User engagement analysis

Clustering based on engagement metrics.

- Optimized number of clusters = 4 based on the elbow plot below.



- Clusters 4 has the highest number of users with an average total sessions duration, average total sessions traffic and an average of 1 xDR session per user.
- Clusters 3 has the 2nd highest number of users with the highest total sessions duration, highest total sessions traffic and an average of 2 xDR sessions per user.
- Clusters 2 has the 2nd lowest number of users with the lowest total sessions duration, average total sessions traffic and an average of 1 xDR sessions per user.
- Clusters 1 has the lowest number of users with an average total sessions duration, average total sessions traffic and an average of 1 xDR sessions per user.

User experience analysis

10 of the top, bottom and most frequent tcp, rtt and tp values in the dataset.

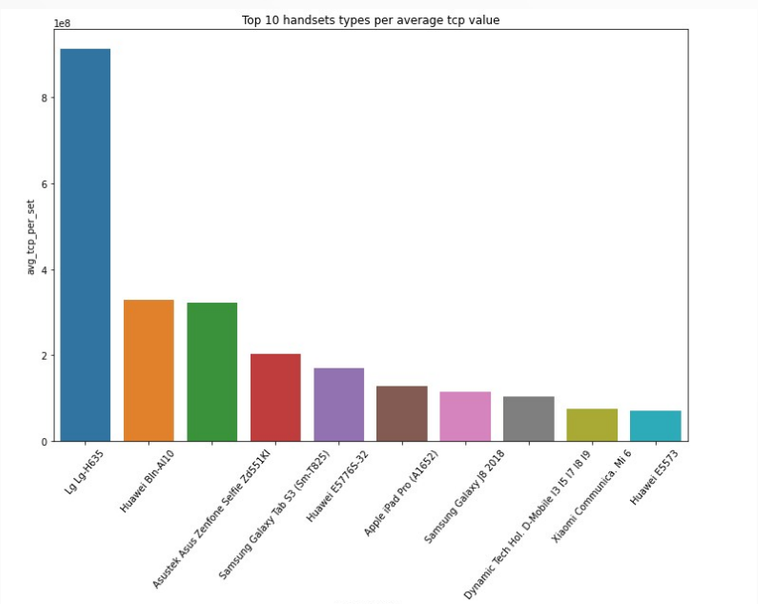
- There’s a variation in the top, bottom and frequent values across the 3 features.
- The variation is due to different units of measurement and the different aspects each feature represents.
- The most frequent values in **RTT** and **TP** are genarally low values unlike frequent **TCP** values.
- The gap between the top and bottom values across the 3 features is extremely large. This can mean the following:
  - If the values are dependent on the handset type, then there are extremely poor performers in the handsets the company stocks.
  - If the values are dependent on user engagement with the applications, then there are extremely less active customers.
- The problem with poor performing devices can be fixed by reducing their stock and increasing the excellent performers. This will result into either forcing users into buying the excellent performers or looking for other companies.
- The less user engagement problem can be fixed by either introducing devices with more applications if it’s an issue of interest or spotting the apps with at least a high engagement among these users and adding more features to them.

User experience analysis

Throuput distribution: The variable is right-skewed unimodal normally distributed with a mean of 12400 and standard deviation of 14661.

Average TCP retransmission per handset type.

- TCP retransmission is an indication of reliable sending of data from end to end.
- The plot below shows that 60% of the handsets with the highest TCP retransmission average values are manufactured by the top 3 handset manufacturers.



User experience analysis

Clusters description on the experience metrics (Total TCP, Total RTT and Total Throughput).

- Cluster 1 has the highest number of users. Thus, on average, most users have 1 type of handset, overally low throughput values, low RTT values and medium TCP values.
- Cluster 2 is the 2nd most populated cluster. On average, users have 2 types of handsets, high RTT values, high throughput values and high TCP values.
- Cluster 3 is the least populated. Users in this cluster have on average: 1 type of handset, low TCP values, medium RTT values and medium throughput values.

Average experience and engagement scores per cluster.

clusters	experience_score	engagement_score
0	5.384672	1.569647
1	9.434669	3.633318

- Cluster 0 has users with lower experience and engagement score compared to cluster 1. As a result, the satisfaction score will be lower in cluster 0.

Limitations, Recommendations and References.

- **Limitations**:
  - Some of the columns in the datasets have large percentages of missing values, thus cannot be analysed.
  - The data has no timestamp. It might contain old records which might give a false overview of the current situation.
  - The analysis is focused only on the user leaving out potential insights that could be obtained from other aspects like handsets-analysis.
  - The call detail record (voice channel) is ignored and more focus is put on data sessions detail record.
- **Recommendations**:
  - The company has some weaknesses and strengths based on the analysis. It is worth buying with some changes to be made:
    - The handset types should be stocked depending on their number of users.
    - More gaming handsets should be introduced given that majority of traffic is as a result of games.
    - More features should be added to the existing gaming apps to increase sessions traffic.
    - Handsets with a wide variety of applications should be introduced to in order to capture the attention of less engaged users.
    - Poor performing handsets based on RTT, TP and TCP values should be replaced by the excellent performers.

- **References**:

  - https://medium.com/@ramdittakavi_61179/exploratory-data-analysis-a-structured-approach-da4d74646445
  - https://blogs.oracle.com/datascience/an-oracle-data-science-case-study-in-telecom
  - https://towardsdatascience.com/how-to-use-python-seaborn-for-exploratory-data-analysis-1a4850f48f14