George Mason University
SYST 699: Masters Capstone Project
Spring 2014

Big Data Analytics on Mobile Usage

May 11, 2014

Arturo Buzzalino
Justin Nguyen
Mitul Patel
Tanner Suttles

TABLE OF CONTENTS

EXECUTIVE SUMMARY

A common pitfall of advertising on mobile devices is a lack of information about the target of the advertising. On mobile this is especially an issue as conversion rate for mobile subscribers is very low. In an effort to provide more targeted and effective advertising SAP is teaming up with mobile phone carriers to provide insights on mobile phone subscribers. Mobile phone carriers have access to a wide variety of data on their customers' utilization of cell phone services, including call, text, web and location data. This data can be anonymized to reduce privacy issues and studied to determine the age and gender of unknown or pre-paid subscribers.

For the Spring 2014 semester the project team was tasked with developing a classification model capable of inferring the gender of a mobile phone subscriber based on their mobile web browsing activity. The team utilized subscriber data from a singular carrier, and produced models using SAP's HANA database with the Predictive Analytics Library (PAL). The project team began by structuring and organizing the data using Extract, Transform, Load (ETL) techniques and used a combination of SQL queries and descriptive statistics to get a big picture of what the data contains. Naive Bayes and CHAID models were developed based on the data to infer the gender of a subscriber.

It was determined that CHAID was more accurate when more parameters and numbers values need to be evaluated; however, Bayes was more accurate with the binary and simplistic inputs in the training set. The algorithms predicted females more than males, but that lead to male predictions being more accurate. Instead of the algorithms having accuracy results based on the demographics of the learning training sets, it turned out the accuracy results were impacted more on the different testing sets. Grouping of the data by age and gender in the training set has little impact on the learning and application of the algorithms on the testing sets.

Data integrity was of high interest to SAP. As this was the first set of data SAP was receiving, the integrity analysis carried out by the team exposed problems within the data that SAP was unaware of. This surprised SAP and they went back to the data providers to go over the issues in detail. SAP requested for higher integrity in the data and was successfully able to communicate this to the data provider using this team's analysis. Being the first data delivery SAP received for Consumer Insight, the team's findings will carry over to other data deliveries and SAP will know what issues to check for to ensure integrity of the data.

The addition of call, text, and location data will expand the amount of data that the current delivered algorithm can learn onto. By adding these parameters, the data would have additional dimensions for differences among genders to become more and more apparent. Logically this exposure of the differences will add more learning to the algorithm, increasing its ability to correctly infer male or female.

Throughout the project, the team learned a great deal about business intelligence, data analytics, and data mining. Although the team was not able to develop a robust enough model to infer gender due to unforeseen circumstances, the team has included a roadmap on the future steps necessary to continue this project to deliver a robust prediction model.

1 INTRODUCTION

1.1 Background

SAP Mobile Services is developing a new product, Consumer Insight 365 (CI365). The purpose of this product is to enhance a business' ability to expand their market, and provide a tool to perform meaningful analysis of consumer patterns. CI365 will analyze large amounts of global mobile carrier data. This mass analysis will extend to a large number of countries across the world and cover millions of people.

The goal of this project is to provide businesses with an additional, powerful means to expand their market by focusing their growth efforts on specific regions and demographics. Data visualization and statistical techniques will be used to determine patterns among: socio-demographics, gender, age, URL click stream categories, geo-location and texting / calling habits. Below is a sample of what a CI365 custom report might look like Figure 1.

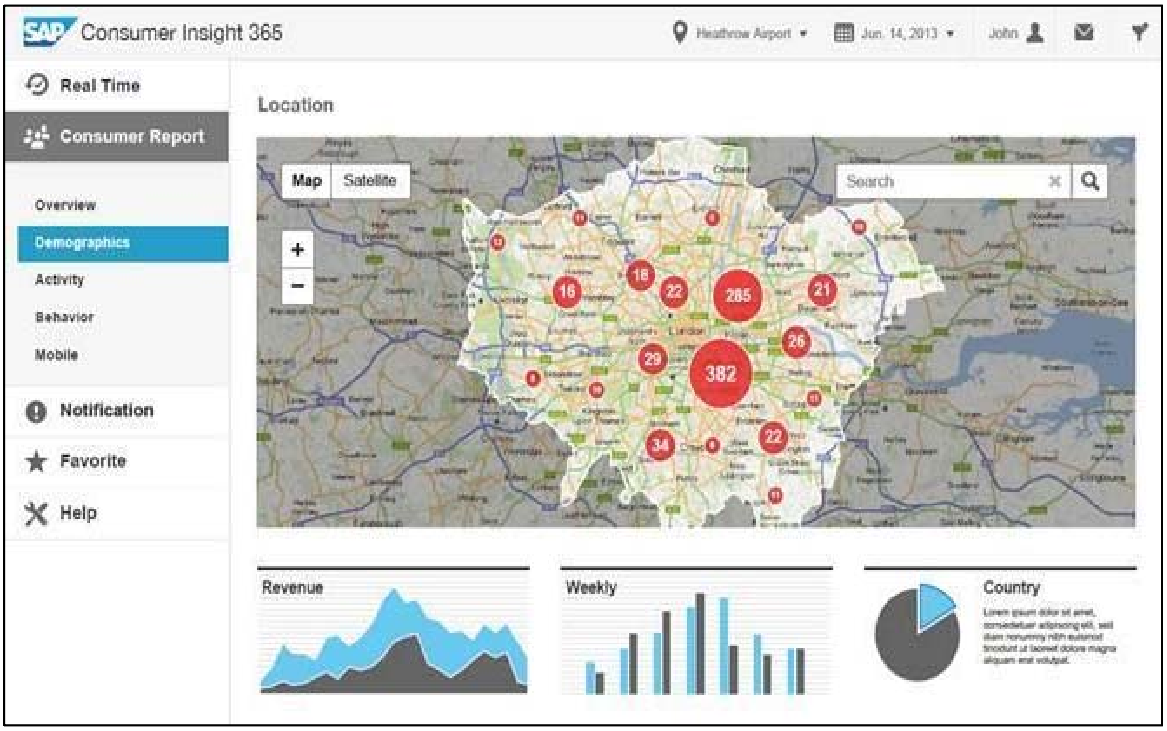


Figure 1: Sample of the type of report a business might be using

The data from the carriers is received in anonymized form, this is important because SAP does not want to breach the privacy of the consumers. There is no way for SAP to trace the number of a person to determine who he or she is or what the address of their home is.

The data provided will include the subscribers' daily data generation through the handset. The age and gender is provided when the subscriber has a plan with the carrier. For those subscribers which are roaming on the network, or own pay as you go and prepaid plans this data is not available. This is where SAP wants to use Big Data analytics to be able to infer a subscriber's gender and age based on their phone habits.

1.2 Problem statement

Focusing on a carrier's mobile user data, determine correlations between texting and calling habits, URL category visits and geo-location with subscriber gender. Using these correlations build a model within SAP HANA PAL to infer the gender of an unknown subscriber.

1.3 Scope

This project is very large and encompasses many different aspects. The project team will be sure to focus on the carrier's data as detailed in the project description. The data from the carrier contains detailed mobile activity from 204,000 subscribers across five days. The team had to reduce the original planned scope due to not receiving certain pieces of data and due to schedule constraints (detailed in section 2.2).

The team will construct a model capable of inferring gender. The model will only consider URL traffic categories and subscriber handsets. The URL traffic categories will be further broken down once into the following:

- # of bytes transferred between website and subscriber
- Start/end times of visiting website
- Top sites visited by gender

To validate the results, the team will test the model on a test set of data and conduct a sensitivity analysis. Upon the completion of the model and testing, the team will submit the following deliverables:

- Data model that implies the gender of the subscriber
- Description and inputs into the model
- Description of patterns that lead to the model
- Result of sensitivity analysis

1.4 Proposal Requirements

Requirement 1: The team shall utilize data provided by mobile carrier
The type of data sent by the mobile carrier includes metadata about user's texting and calling habits, points of interest frequented (geo-location), and URLs visited

Requirement 2: The team shall develop methods to identify patterns in cell phone usage by gender
The methods used will be developed based on statistical and data-mining principles and techniques. These methods shall consistently identify pattern in cell phone usage by gender.

Requirement 3: The team shall develop methods to identify patterns in cell phone usage by age group
The methods used will be developed based on statistical and data-mining principles and techniques. These methods shall consistently identify patterns in cell phone usage by age group.

Requirement 4: The team shall develop a model for classifying a subscriber's gender
The team will develop a model within SAP HANA for classifying the gender of a subscriber.

Requirement 5: The model shall predict the gender of an anonymized user as male or female
The model will be developed based on patterns identified to predict the gender of a user.

Requirement 6: The model shall predict the age group of an anonymized user
The model will be developed based patterns identified techniques to predict the age group of a user.

Requirement 7: The model shall provide accuracy for each classification
The model will produce accuracy of its classification result for each subscriber.

2 TECHNICAL APPROACH

2.1 Data Description and Terminology

SAP performed a variety of operations in order to anonymize the data before providing it to the team. The subscriber data was compiled into a single table consisting of a record for each subscriber's transaction within the mobile network. For web browsing this equated to one record for every URL request the subscriber made. A simplified example of the data is shown below in Table 1.

Table 1: Example Data

RECORD ID	CARRIER ID	SUB ID	START TIME	END TIME	BYTES IN	BYTES OUT	DOMAIN	AGE BAND	GEN DER	HANDSET	ZIP CODE
798,381,591	42	1-FXO-689	1/24/14 1:10	1/24/14 1:10	0	7,320	'meta.radioactive.sg'	18-24	M	Samsung N7105	259
798,063,689	42	1-FXO-689	1/24/14 1:10	1/24/14 1:10	0	7,320	'meta.radioactive.sg'	18-24	M	Samsung N7105	259
798,296,458	42	1-FXO-689	1/24/14 1:10	1/24/14 1:10	0	7,320	'meta.radioactive.sg'	18-24	M	Samsung N7105	259
798,353,256	42	1-FXO-689	1/24/14 1:10	1/24/14 1:10	0	7,320	'meta.radioactive.sg'	18-24	M	Samsung N7105	259

A feature of the web browsing data that complicates its use is that a single browsing experience can manifest in a variety of ways in the transactional data. A record is saved for each URL request. For conventional web browsing this equates to every time a page is clicked or something is type into the search/URL area of the browser. For streaming video and audio the URL is requested without the subscriber's direct interaction, averaging 15 transactions per second. When trying to determine how long a subscriber is viewing a web page people who read news articles online may only have transactions for a second, but read the article for several minutes versus an online radio listener who has a transaction for every second they are on the service. Below, Table 2 outlines the details of the data received.

Table 2: Field Descriptions

Field Name	Description and Use
RECORD_ID	A unique ID for every transaction in the database.
CARRIER_ID	A coded number for the carrier.
SUB_ID	The Subscriber ID is a unique value that describes the account and phone line that the record came from. Every cell phone subscriber is assigned an ID.
START_TIME	The starting time for the transaction.
END_TIME	The ending time for the transaction. The team found for all data transactions the end time matched the starting time.
BYTES_IN	The number of bytes received by the network.

BYTES_OUT	The number of bytes received by the mobile device.
DOMAIN	The URL Domain requested by the subscriber.
CATEGORY	The categorized domain.
AGE_BAND	The age band of the subscriber, defined as the groups: ● <18 ● 18-24 ● 25-34 ● 35-44 ● 45-54 ● 55-59 ● 60+
GENDER	The gender of the subscriber
HANDSET	The handset the subscriber was using for this transaction. The team found that some subscribers switch handsets during the time period the data sampled. It was also found that some subscribers switched handsets 50+ times, which indicated corrupt data.
HOME_POST_CODE	The local billing post code for the subscriber.

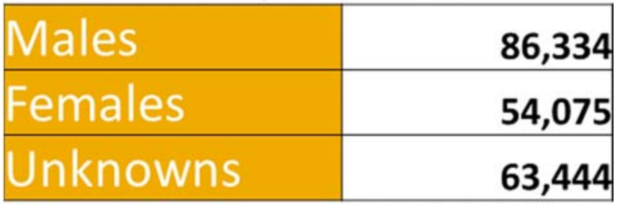
2.2 Analysis

The team had to get a high level understanding of what type of information was in the data. The team achieved this by manipulating the data with a series of SQL queries to derive meaningful statistics from the data set. Information captured includes:

Demographics:

- Number of subscribers whose age group and/or gender is known
- Number of subscribers whose age group and/or gender is unknown
- Total number of subscribers by gender and age group

Table 3: Subscribers by Gender



Males	86,334
Females	54,075
Unknowns	63,444

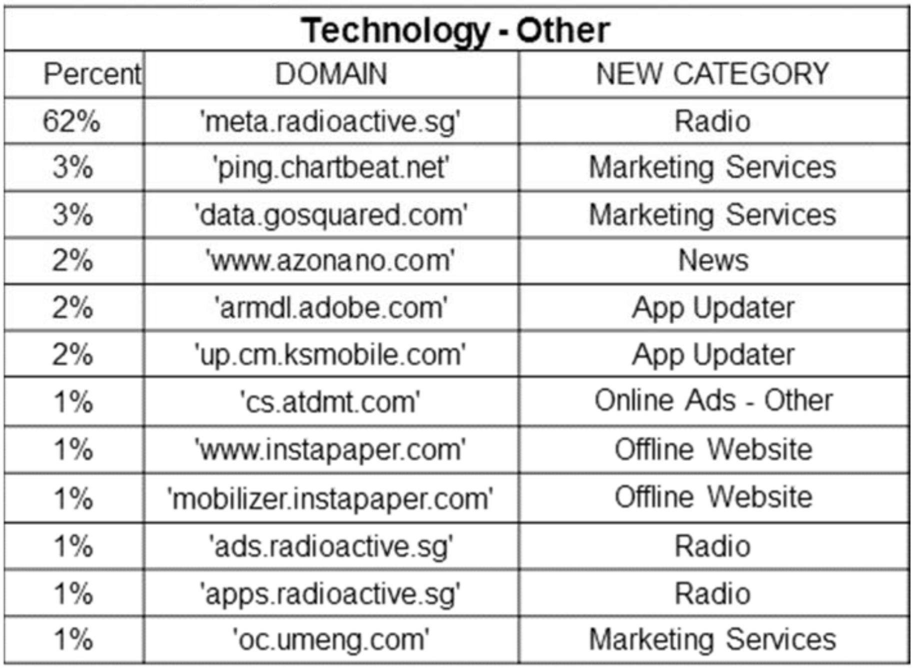
Attributes By Gender:

- Top URL categories visited
- Number of subscriber domain activities per hour per day
- Number of subscribers per hour per day
- Top handsets used
- Home zip code

The data included all of the URL's that subscribers visited. Since there are so many unique URL's, the team used a categorization service to categorize the URL's. For example, URL's to Amazon, eBay, and Craigslist would be categorized as Shopping. This insight into the data gave the team a foundation to determine the appropriate data mining algorithms to use.

When the URLs were categorized using the category API tool, 2 categories in particular (Technology – Other and Uncategorized) encompassed most of the URLs in the data. The team identified the domains that covered 80% of the traffic in each category and assigned new categories. Table 4 below, shows the mapping and re-categorization for the Technology – Other category.

Table 4: Re-categorizing Domains



Technology - Other

Percent	DOMAIN	NEW CATEGORY
62%	'meta.radioactive.sg'	Radio
3%	'ping.chartbeat.net'	Marketing Services
3%	'data.gosquared.com'	Marketing Services
2%	'www.azonano.com'	News
2%	'armdl.adobe.com'	App Updater
2%	'up.cm.ksmobile.com'	App Updater
1%	'cs.atdmt.com'	Online Ads - Other
1%	'www.instapaper.com'	Offline Website
1%	'mobilizer.instapaper.com'	Offline Website
1%	'ads.radioactive.sg'	Radio
1%	'apps.radioactive.sg'	Radio
1%	'oc.umeng.com'	Marketing Services

Similarly to the URL categories, the handsets were irregular and too specific in naming. For example there were 57 different Samsung handsets, and the team mapped this to a single Samsung category. This was done for all the others.

Another method the team used to analyze the data was to attempt to determine the duration a subscriber spent browsing within a category. The web data transactions did not contain a duration field, so there was no way to easily tell how long the user was spending on a website. To approximate the duration the data was grouped into five minute intervals. If a subscriber visited a category within that five minute interval the subscriber was credited with five minutes of viewing time. In addition to these durations the count of transactions within the five minute span was totaled as the subscribers' activity in that period.

The team recognized multiple anomalies within the data sets that would have to be filtered out in the model data sets. There was a large amount of transactions with null domains which was unusable for the model. The amount of subscriber domain activities shows drastic and random peaks and lows per hour per day. The data set is also missing hours in certain days of data. This is shown in Figure 3 where the number of subscriber and activity count per hour is graphed for the 5 days.

The team originally looked at using tools such as R, Weka, and SAP High-Performance Analytic Appliance (HANA) Predictive Analysis Library (PAL). The team went with SAP HANA PAL since the data was stored on the SAP HANA server and it would be extremely impractical to port the data over to Weka or R due to its sheer size (over 500GB). SAP HANA PAL contains all the statistical and data mining algorithms required for the project and are robust enough to analyze the large amount of data available.

2.3 Data Patterns

The first step in the data analysis was to develop some overall views of the data to determine the quality. The first plot generated was a distribution of the subscribers by age and gender, shown below in Figure 2. This distribution followed the team's assumptions that the bulk of the subscribers would be concentrated in the 25-54 age range.

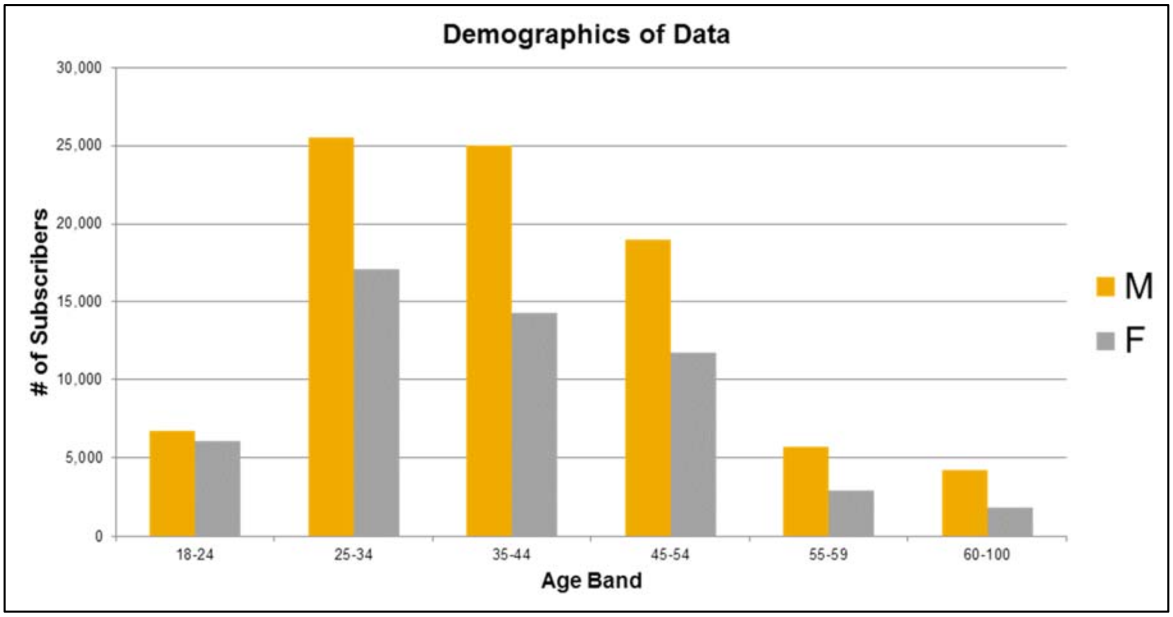


Figure 2: Population Demographics

After running across some strange transaction counts for particular subscribers, a chart was generated showing the number of distinct subscribers and the total activity for each hour of the dataset. This plot is shown in Figure 3 and demonstrates a few data quality issues that were encountered. On the first day there are a roughly equivalent number of subscribers as the other days; however the activity for that day is almost zero. There were also several spikes in the number of transactions that could be traced to single individuals sending and receiving terabytes of data in a single hour. These subscribers were removed from the analysis as their data was either corrupt or their activity was outside the boundaries of normal use. There were also several instances where an hour of data would be missing from the dataset, including a span on the 24th and 26th where 14 hours of data were missing from the set. Lastly from a qualitative perspective the plots of the subscribers and activity seem highly irregular. The team assumed the amount of activity would follow a pattern day to day with heavier usage during awake hours. With the high number of users, roughly 200,000, there should be much higher stability and an obvious pattern in the data generated and for that SAP was made aware that there were problems with the data provided.

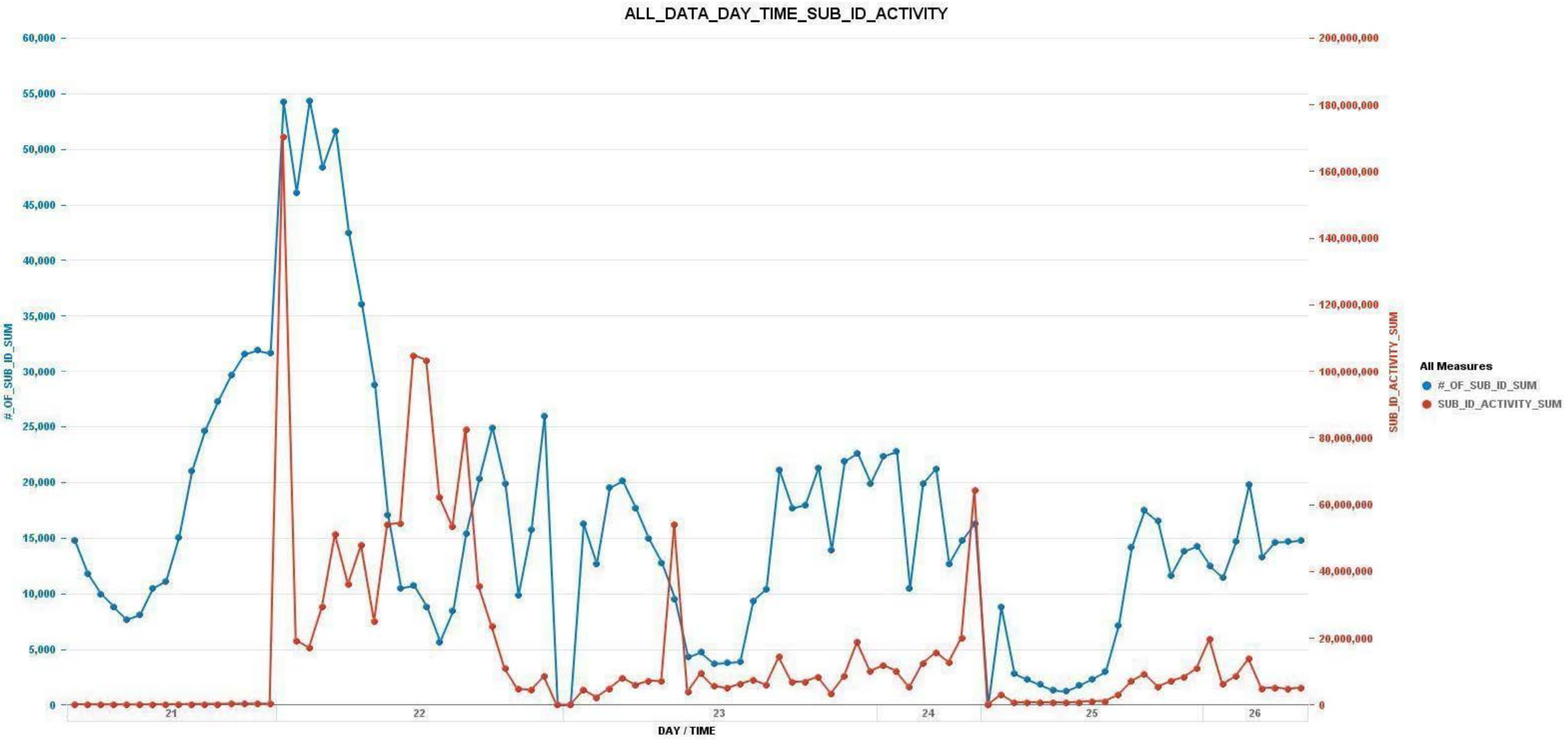
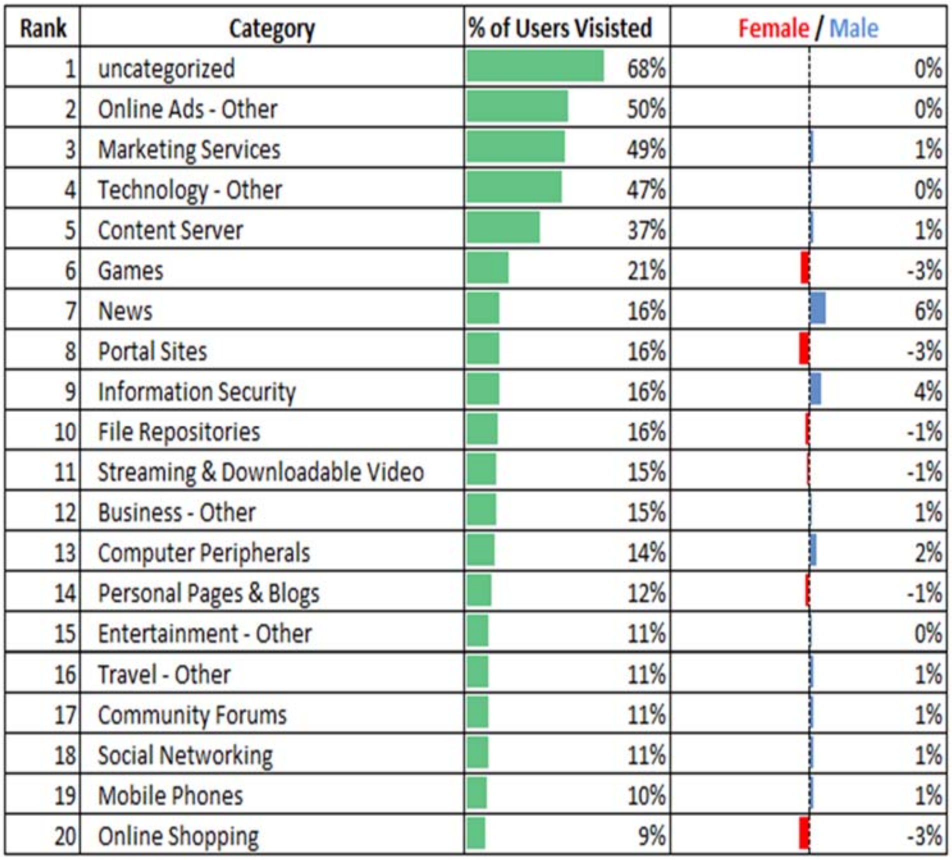


Figure 3: Subscribers and Activity

The following table shows the top visited categories by the training data set. The columns are the category's rank, description, the percent of the subscribers that visited and the gender bias of the number of subscribers who visited. The bias is shown as the deviation from a 50/50 split. The table of the top categories demonstrates how quickly the visitation drops off in the categories and how little difference there is in the visitation by males and females. With the top 20 categories there is very little to differentiate between genders in category visitation.

Table 5: Top 20 Categories



Rank	Category	% of Users Visisted	Female / Male
1	uncategorized	68%	0%
2	Online Ads - Other	50%	0%
3	Marketing Services	49%	1%
4	Technology - Other	47%	0%
5	Content Server	37%	1%
6	Games	21%	-3%
7	News	16%	6%
8	Portal Sites	16%	-3%
9	Information Security	16%	4%
10	File Repositories	16%	-1%
11	Streaming & Downloadable Video	15%	-1%
12	Business - Other	15%	1%
13	Computer Peripherals	14%	2%
14	Personal Pages & Blogs	12%	-1%
15	Entertainment - Other	11%	0%
16	Travel - Other	11%	1%
17	Community Forums	11%	1%
18	Social Networking	11%	1%
19	Mobile Phones	10%	1%
20	Online Shopping	9%	-3%

Box plots were also generated for the activity of the subscribers within the category. The box plot for the Technology Other category is shown in Figure 4. The box plots in the top 20 categories look similar in that there is very little differentiation between how the two genders visited the category. Both genders were roughly equivalent in their number of visits to a category. The remainder of the box plots are included in Appendix D. The red and green coloring of the box plots were a function of the graphics software used to plot the data, and do not hold any significance other than differentiating the middle quartiles.

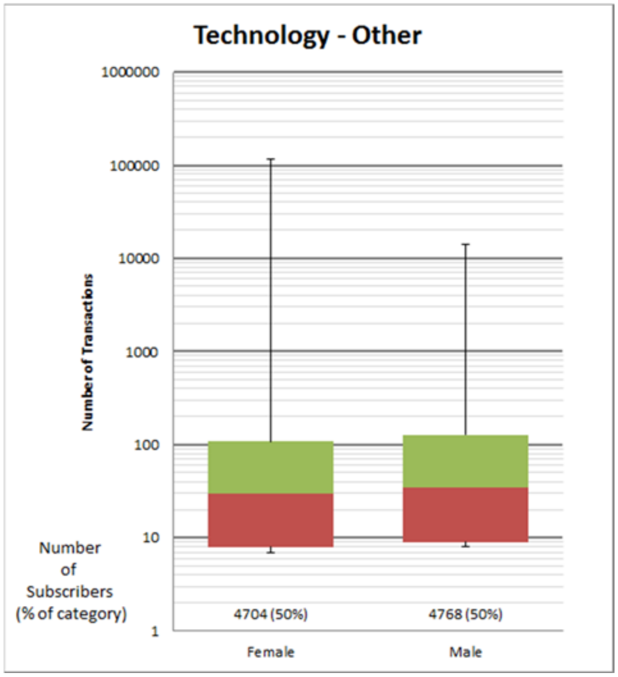
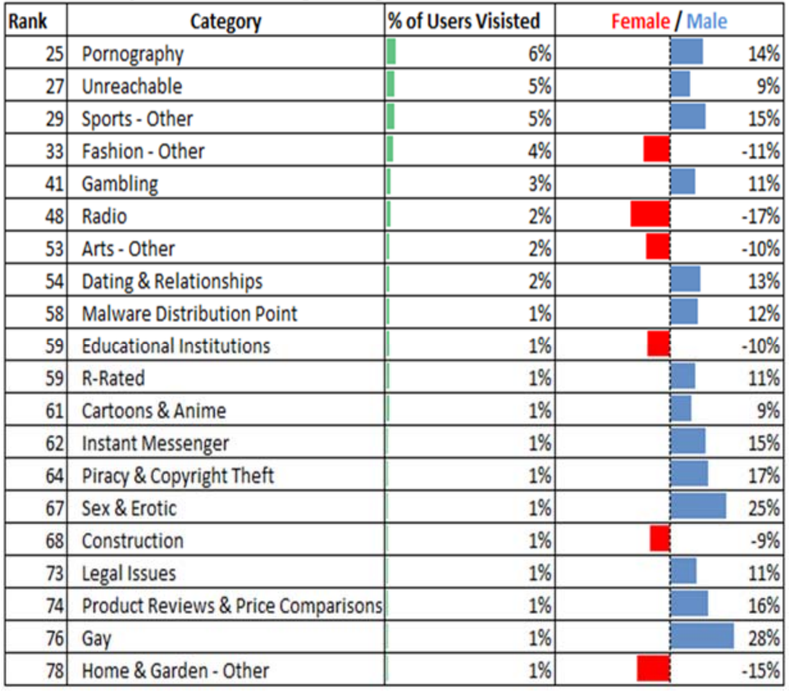


Figure 4: Technology Other Transaction Count Box Plot

Moving farther down the list of top categories there starts to be differentiation between the number of male and female visitors to the category. Table 6 is a filtered list of categories. The list is limited to categories with visits from more than 1% of the subscribers and more than a 16% swing in the gender that is a split greater than 58%/42%. With this list there are good categories for determining the gender of the subscriber, however the visitation of these categories is quite low and the chances of an unknown subscriber visiting one of these categories is low. Given that these categories were the strongest differentiators that were found, a model was developed using only these categories.

Table 6: Top Differentiating Categories



Rank	Category	% of Users Visisted	Female / Male
25	Pornography	6%	14%
27	Unreachable	5%	9%
29	Sports - Other	5%	15%
33	Fashion - Other	4%	-11%
41	Gambling	3%	11%
48	Radio	2%	-17%
53	Arts - Other	2%	-10%
54	Dating & Relationships	2%	13%
58	Malware Distribution Point	1%	12%
59	Educational Institutions	1%	-10%
59	R-Rated	1%	11%
61	Cartoons & Anime	1%	9%
62	Instant Messenger	1%	15%
64	Piracy & Copyright Theft	1%	17%
67	Sex & Erotic	1%	25%
68	Construction	1%	-9%
73	Legal Issues	1%	11%
74	Product Reviews & Price Comparisons	1%	16%
76	Gay	1%	28%
78	Home & Garden - Other	1%	-15%

Similarly to the top 20 categories, box plots of the subscriber activity in the distinguishing categories revealed little difference in how much activity each gender had in the category. Plots of the Sports and Fashion categories have been included below in Figure 5 and Figure 6. The Sports category visitation was biased towards male, while the fashion category visitation was biased towards females. There are some distinctions in the activity; the biased gender has a slightly higher amount of activity with greater maximum and minimum activity values. The middle 50% of each gender still overlaps almost entirely, which makes the amount of activity in these categories a poor classifying attribute.

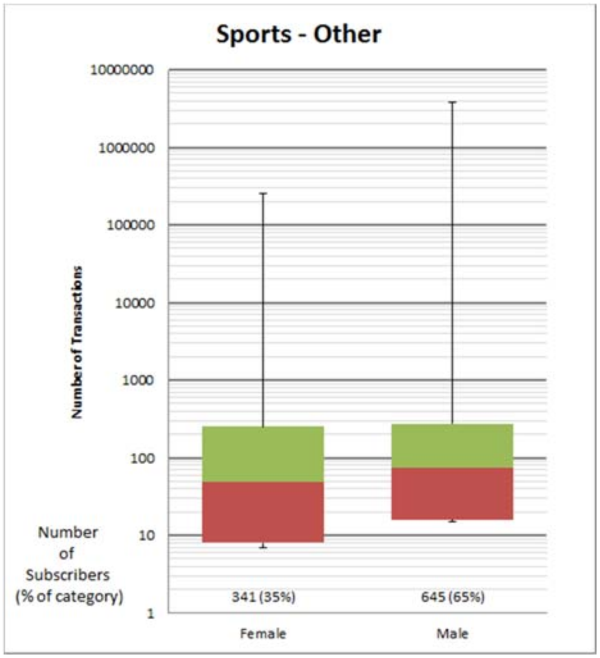


Figure 5: Sports Other Transaction Count Box Plot

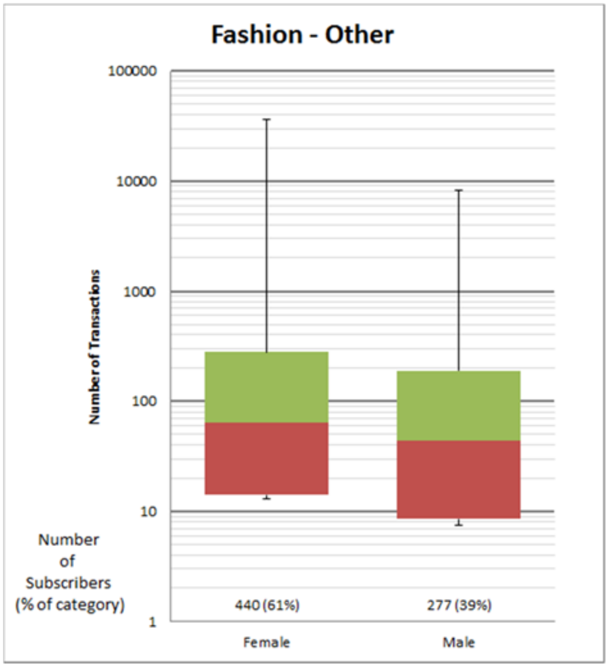


Figure 6: Fashion Other Transaction Count Box Plot

2.4 Model Development

Based on a variety of literature reviews and consultation with advisors, the team decided to focus on two classification algorithms: Naïve Bayes and Chi-Squared Automatic Interaction Detection (CHAID). Running both classification algorithms on the dataset will help the team validate and provide a baseline to evaluate the results.

Naive Bayes is simple in allowing each attribute to contribute towards the final decision equally and independently from the other attributes. This simplicity equates to computational efficiency, which makes NB techniques attractive and suitable for many domains. The Naive Bayes algorithm is based on conditional probabilities. It uses Bayes' Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data.

CHAID constructs non-binary trees for classification problems (when the dependent variable is categorical in nature) relies on the Chi-squared test to determine the best next split at each step. It prepares predictions by creating categorical predictors out of continuous predictors. It cycles through the predictors to determine for each predictor the pair of categories that is least significantly different with respect to the dependent variable by computing a Chi-square test. If the test shows a pair of predictors is not statistically significant, the predictor categories will merge.

There are two sets of data that need to be provided to these classifications algorithms, the training and the testing data sets. The training set will have the correlation parameters and the class attributes (gender) to be inputted into the algorithm for it to learn and generate a trained result. That trained result is inputted into the testing set where the data will just have the correlation parameters without the gender and it will generate a gender prediction. The testing set will have different subscribers from the training set and the gender will be known so it can be evaluated with the predicted results from the model.

The team used the software, SAP HANA Studio, to create and run all the sequel scripts to create the training and testing sets from the data. Due to the million rows of data, SAP Data Services Designer was used pivot certain rows to columns when needed. SAP HANA PAL was used to create models that ran the Naïve Bayes and CHAID algorithms on the training and testing sets. Unfortunetly this software did not provide any confidence with the results. The processs and scripts of generating the data sets and running them on the algorithms is shown in Appendix C.

2.5 Model Evaluation

An evaluation of whether the data model’s gender implication is accurate shall be performed. Two forms of testing conducted will be testing of accuracy and performing a sensitivity analysis. The model’s accuracy is measured by the number of correct predictions of gender and age group. The sensitivity analysis will vary the input metrics of the data model and see how the output results change. Many forms of training and testing sets will be developed and evaluated to determine the best method to get accurate results for both classification algorithms. This will reduce the amount of uncertainty in the model as well as increase the understanding of relationships between the input variables and output result. The testing stage will also ensure that all the requirements are being fulfilled by the data model.

3 RESULTS

3.1 Phase 1 – Raw Transactions

The training set had separate rows of same subscribers due to same subscriber going to multiple URL categories. This can be seen in Table 7. The classification algorithms would evaluate just the category activities and apply it to the testing set. The problem with this training set is that the results would have different gender results for the same subscriber. This training and testing set is cannot be evaluated and needs to be organized in a manner that the algorithm can provide one gender result per subscriber. Accuracy could not be calculated for Phase 1 due to multiple gender results for single distinct subscriber.

Data Set Definitions
Training Set 1 = 10,000 distinct subscribers for each gender w/ random age bands (20k total)
Testing Set 1 = 500 distinct subscribers for each gender w/ random age bands (1k total)

Table 7: Phase 1 - Training Set Example

SUB_ID	CATEGORY	ACTIVITIES	GENDER
1	Sports	3245	M
1	Gambling	234	M
1	News	87	M
2	Shopping	1123	F
2	News	853	F
3	Technology	2	M
3	Games	769	M

3.2 Phase 2 – Pivoted Category with Activities and Duration Span

The training and testing set 1 was pivoted with 150 categories with activity and duration span values (Detailed in section 3.2). With the pivot approach the algorithm would only generate one gender result per subscriber. The team used the SAP software Data Services Designer to pivot the million plus rows of data. The classification algorithms would also evaluate the subscriber's home zip, handset, age band, category activities, and category duration span. Even though the unknown data set will not have age band, the team wanted to evaluate how accurate the results would be with it. An example of the pivoted table layout is shown in Table 8. The "..." column represents the remaining 150 categories with the activities and duration span per distinct subscriber. When certain subscribers do not visit certain URL categories, it places a '0' in the activities and duration span cells. Table 9 shows the results of phase 2.

Data Set Definitions
Training Set 1 = 10,000 distinct subscribers for each gender without defining age bands (20k total)
Testing Set 1 = 500 distinct subscribers for each gender without defining age bands (1k total)

Table 8: Phase 2 - Training Set Example

SUB_ID	HOME_ZIP	AGE_BAND	HANDSET	ART_ACTIVITIES	ART_DURATION_SPAN	...	GENDER
1	710	25-34	Sony	0	0	...	M
2	540	25-34	Samsung	1763	15	...	F
3	679	35-44	Apple	0	0	...	M

Table 9: Phase 2 - Results

Phase	Train Set	Test Set	Algorithm	Total Accuracy
2	1	1	Bayes	50%
2	1	1	CHAID	55%

3.3 Phase 3 – Pivoted Category with Binary Activities

In this phase a binary approach was taken; the algorithm would evaluate whether the subscriber visited a URL category or not. When a subscriber visits a certain URL category; the cell value will be ‘1’. When the subscriber does not visit a certain URL category, the cell value will be ‘0’. The purpose of this approach is to see if the algorithm prefers simple inputs rather than reducing its affinity with category activities and duration span values. To keep its simplicity, taking a bottom up approach, the home zip, handset, and age band were removed from the training and testing sets. So the training and testing sets for this phase had the 150 URL categories in binary form and gender per subscriber. An example of this approach is shown in Table 11. The “...” column represents the remaining 150 categories with the binary activity values per distinct subscriber. Table 12 shows the results of phase 3.

Data Set Definitions
Training Set 1 = 10,000 distinct subscribers for each gender without defining age bands (20k total)
Testing Set 1 = 500 distinct subscribers for each gender without defining age bands (1k total)

Training Set 2 = 20% of all distinct age band of all subscribers for each gender (14k M, 14k F, and 28k total) shown in Table 10.
Testing Set 2 = 3500 distinct subscribers for each gender without defining age bands (7k total)

Table 10: Training Set 2 Subscriber Distributions per Age Band

Age Band	Male	Female
18-24	700	700
24-34	2150	2150
35-44	2000	2000
45-54	1500	1500
55-59	450	450
60-100	300	300

Table 11: Phase 3, 4, and 5 - Training Set Example

SUB_ID	ART_ACTIVITIES	NEWS_ACTIVITIES	...	GENDER
1	0	1	...	M
2	1	0	...	F
3	0	1	...	M

Table 12: Phase 3 Results

Phase	Train Set	Test Set	Algorithm	Total Accuracy
3	1	1	Bayes	62%
3	1	1	CHAID	50%
3	2	2	Bayes	55%
3	2	2	CHAID	50%

3.4 Phase 4 – Master Test Set for Three Training Sets

The purpose of this phase was to evaluate how the three training sets in both algorithms perform on a single testing set. If a significant improvement in results is observed between training sets, it can be inferred that the demographic grouping techniques for that training set give the best result. It was decided to create a single testing set to evaluate each algorithm. This master test set does not have any of the same distinct subscribers from the previous training sets. The approach of pivoting 150 URL category binary activities was kept because a satisfactory result was achieved in the last phase with this method. This is shown in Table 11. The "..." column represents the remaining 150 categories with the binary activity values per distinct subscriber. Table 13 shows the results of phase 4.

Data Set Definitions
Training Set 1 = 10,000 distinct subscribers for each gender without defining age bands (20k total)
Training Set 2 = 20% of all distinct age bands of all subscribers for each gender (14k M, 14k F, and 28k total).
Training Set 3 = 1000 distinct subscribers per age band per gender (6k M, 6k F, and 12k total)

Master Testing Set = 500 distinct subscribers per age band per gender (3k M, 3k F, and 6k total)

Table 13: Phase 4 Results

Phase	Train Set	Test Set	Algorithm	Male Accuracy	Female Accuracy	Total Accuracy
4	1	Master	Bayes	56%	53%	54%
4	1	Master	CHAID	52%	52%	52%
4	2	Master	Bayes	57%	54%	55%
4	2	Master	CHAID	52%	51%	52%
4	3	Master	Bayes	58%	54%	55%
4	3	Master	CHAID	51%	52%	52%

3.5 Phase 5 – Subsampling Categories

The purpose of this phase was to only have subscribers that used the top 27 URL categories in the training and testing sets (outline in Table 6 section 3.3). Instead of pivoting 150 categories, there is only 27 categories that needed to be pivoted. The purpose of this is to see if the algorithms will detect the 16% and higher gender difference in categories. The team is also examining two separate testing sets (Testing set 1 and the Master test set) with one training set (Training set 2) and seeing if there are any significant differences. The distinct subscribers that only visited the 27 categories were used; this caused the number of distinct subscribers to change from its original value of 28,000 to 10,000 for the training set and the Master testing from 6,000 to 3,200. The URL category binary activities method was still used in this phase. This is shown in Table 11. The "..." column represents the remaining 27 categories with the binary activity values per distinct subscriber. Table 14 shows the results of phase 5.

Data Set Definitions
Training Set 2 = 10,000 total distinct subscribers with only top 27 gender differentiating categories
- (Removal of subscribers outside of categories reduced the total from 28k to 10k)

Testing Set 1 = 1000 total distinct subscribers with only top 27 gender differentiating categories
- (same amount of subscribers from original set)

Master Testing Set = 3,200 total distinct subscribers with only top 27 gender differentiating categories
- (Removal of subscribers outside of categories reduced the total from 6k to 3.2k)

Table 14: Phase 5 Results

Phase	Train Set	Test Set	Algorithm	Total Accuracy
5	2	1	Bayes	62%
5	2	1	CHAID	50%
5	2	Master	Bayes	55%
5	2	Master	CHAID	52%

3.6 Discussion

Phase 1 had no results because the tables were arranged in a manner where multiple gender predictions were occurring for a single distinct subscriber. After pivoting the tables to fix this issue for Phase 2, the team realized that Bayes predicted all females while CHAID had better male and female distribution predictions. However, the accuracy of the CHAID results was only 45%. Since Bayes was not providing significant results, the team took a simpler approach and let the algorithm learn whether or not the subscriber visited certain URL categories.

With this new binary approach, Phase 3 shows that the Bayes algorithm is not predicting all females anymore and has increased in accuracy up to 62% for one of the training and testing sets. Also, in this binary approach Bayes has a much better performance than CHAID. Bayes had an accuracy of 62% and 55% while CHAID showed an accuracy of 50% for testing sets 1 and 2. This pointed to the fact that CHAID works better than Bayes when more parameters and number values are inputted into the training set.

Phase 4 used three different training sets that were grouped differently by age band and gender to see how the algorithms would react to one master testing set. The testing set does not have any same subscribers as the training sets. Even though Bayes still provided better results than CHAID, the results for all three training sets were the same. The Bayes accuracy was 55% and the CHAID accuracy was 52% for all three training sets. It seemed like the demographics of the training sets had no impact on the results. This phase also shows that the accuracy of the male predictions was higher because it predicted females more often.

Phase 5 narrowed the URL categories down to the top 27 gender differentiating categories that showed more than a 16% difference in gender usage (Table 6 section 3.3). This time only one data set was used to train the model and it was evaluated against two testing sets. Narrowing the categories made no significant difference in accuracy resulting from the testing sets. However, the algorithms had different results, similar to the results the testing sets received in Phases 3 and 4.

Overall, CHAID is more accurate when more parameters and number values need to be evaluated; however, Bayes is more accurate with the binary and simplistic inputs in the training set. In most phases, the algorithms predicted females more than males, but that lead to male predictions being more accurate. Instead of the algorithms having accuracy results based on the demographics of the learning training sets, it turned out the accuracy results were impacted more on the different testing sets. Grouping of the data by age and gender in the training set has little impact on the learning and application of the algorithms on the testing sets. It was also unfortunate that HANA PAL did not provide confidence results with the Naïve Bayes accuracy results.

4 CONCLUSIONS

As outlined in the Scope section, the roadblocks forced the team to redefine the deliverables. The main delivery change is a model capable of inferring only gender and the exclusion of an age inferring algorithm. Due to the steep learning curve using the HANA tools, the late data delivery and issues with data integrity, the team limited the model generation to inferring gender. Along with the model, the delivery includes details of generating the model, its accuracy and sensitivity under varying testing scenarios. All of the requirements of the project proposal were met with the exception of creating a model that classifies subscribers by age band.

The gaps in the data limited the team's analysis greatly, as the only data the team was able to use were home zip codes, URLs and handset type in order to infer gender. Coupled with the late delivery of the data, the team was heavily limited in the depth of the analysis. Despite these setbacks the recommendations delivered were beneficial to SAP.

Data integrity, detailed in sections 3.2 and 3.3, was of high interest to SAP. As this was the first set of data SAP was receiving, the integrity analysis carried out by the team exposed problems within the data that SAP was unaware of. This surprised SAP and they went back to the data providers to go over the issues in detail. SAP requested for higher integrity in the data and was successfully able to communicate this to the data provider using this team's analysis. Being the first data delivery SAP received for Consumer Insight, the team's findings will carry over to other data deliveries and SAP will know what issues to check for to ensure integrity of the data.

The team was able to design a model capable of identifying a Male or Female with only URL categories with accuracy slightly better than chance. Considering the limited time and faulty data, the team believes that approach can be a strong basis for other models that have more complete data sets. The binary category decisions (outlined in section 4.3), couple with a more full data set where call durations, texts, and geo-location are known, can yield a model with much higher accuracy. The delivery of the team's model will facilitate further analysis

5 RECOMMENDATIONS AND FUTURE WORK

Along with the delivered model, the team suggests SAP take additional steps to expand the findings. The immediate next step is for SAP to run the model delivered on the new one month data. This new set is complete, and does not have the data issues the team faced with this project. There is promise that the delivered model will have much higher efficacy with this complete set and yield much better results. The team also advises the following additional research.

Instead of using the categorization API on the domain alone, this classification algorithm should be run on the full URN path of the websites. This will add greater granularity to the data and expose greater difference amongst genders. To put this into context, in the current methodology cnn.com/basketball and cnn.com/finance have the same domain, cnn.com. In this instance both fall under “News”, but when taking URN path into account they suddenly differ and they become “Sports - Basketball” and “Finance” respectively. Adding this granularity expands the possibilities for there to be differences among the genders.

The addition of call, text, and location data will expand the amount of data that the current delivered algorithm can learn onto. By adding these parameters, the data would have additional dimensions for differences among genders to become more and more apparent. Logically this exposure of the differences will add more learning to the algorithm, increasing its ability to correctly infer male or female.

SAP should work internally and find other methods of developing algorithms and not limit the analysis to PAL (Predictive Analytics Library). SAP should work with data experts both internally and externally with 3rd party partners to determine the best methodology to employ when developing these learning models. Working directly with the PAL team SMEs, SMEs in R-statistics for Hadoop and 3rd party analytics providers, SAP can greatly improve the accuracy of the model provided

Apply the same steps outlined for gender prediction, in order to infer the age-band of the subscriber
With the delivered model, along with the recommendations provided, SAP will have the ability to employ models capable of determining age and gender with a much greater ability than the current system. With this ability Consumer Insight will have the necessary applicability to give a full picture of the consumers being analyzed and not fall behind when the age and gender of the subscriber is unknown.

6 APPENDIX A: PROJECT MANAGEMENT APPROACH

6.1 Project Plan

The project was divided into four task areas: project management, research, model development, and final deliverables. The timeline shown in Figure 7 shows the research phase in blue, model development in green and final deliverables in yellow. The project management task covers initial project activities like kickoff and problem definition, along with progress presentations throughout the project. The introduction of new tools and concepts for the team in the area of big data also required a period of research in which the team learned big data analytic techniques and current cell phone usage research. The majority of the project was spent in data analysis and model development. Near the end of the semester the team started to look toward the final deliverables, and time spent on model development was shared with drafting the presentation. The majority of the time for the final deliverables was spent on the final presentation. The work on the final presentation started earlier in the project with professor reviews, to help organize the information for the final report. At the end of the project the final report and presentation were delivered and the project closed out.

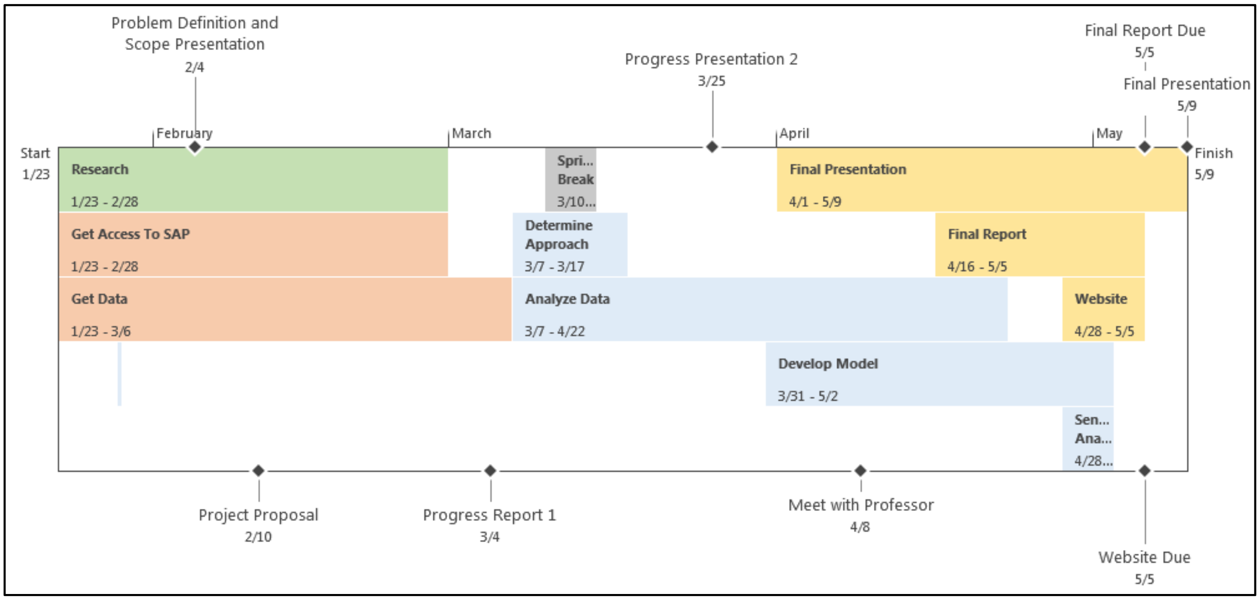


Figure 7: Project Timeline

An abbreviated work breakdown structure (WBS) has been included below in Table 15. The abbreviated version shows major tasks and milestones in the project schedule. The full WBS is included in Appendix B.

Table 15: Summarized project Work Breakdown Structure

Outline Number	Task Name	Start	Finish	Duration
1	**Project Initiation**	**23-Jan**	**25-Mar**	**44 days**
2	**Research**	**23-Jan**	**28-Feb**	**27 days**
2.1	Mobile Phone Use Demographics	23-Jan	14-Feb	17 days
2.2	Big Data Tools	23-Jan	28-Feb	27 days
3	**Model Development**	**23-Jan**	**2-May**	**72 days**
3.1	Get Access To SAP	23-Jan	28-Feb	27 days

3.2	Get Data	23-Jan	6-Mar	31 days
3.3	Determine Approach	7-Mar	17-Mar	7 days
3.4	Analyze Data	7-Mar	22-Apr	33 days
3.5	Develop Model	31-Mar	2-May	25 days
3.6	Sensitivity Analysis	28-Apr	2-May	5 days
4	**Website**	**28-Apr**	**5-May**	**6 days**
5	**Final Report**	**16-Apr**	**5-May**	**14 days**
6	**Final Presentation**	**1-Apr**	**9-May**	**29 days**

6.2 Project Reporting

The team tracked their hours each week using a spreadsheet, which was then used as the basis of determining earned value reporting metrics. Figure 8 shows the reporting of Planned Value (PV), Earned Value (EV) and Actual Cost (AC). Planned value was determined by all team members working ten hours a week on the project. The chart shows the project falling behind schedule through the middle of the semester and then catching back up at the end, this is detailed in the following chart summarizing schedule and cost variance.

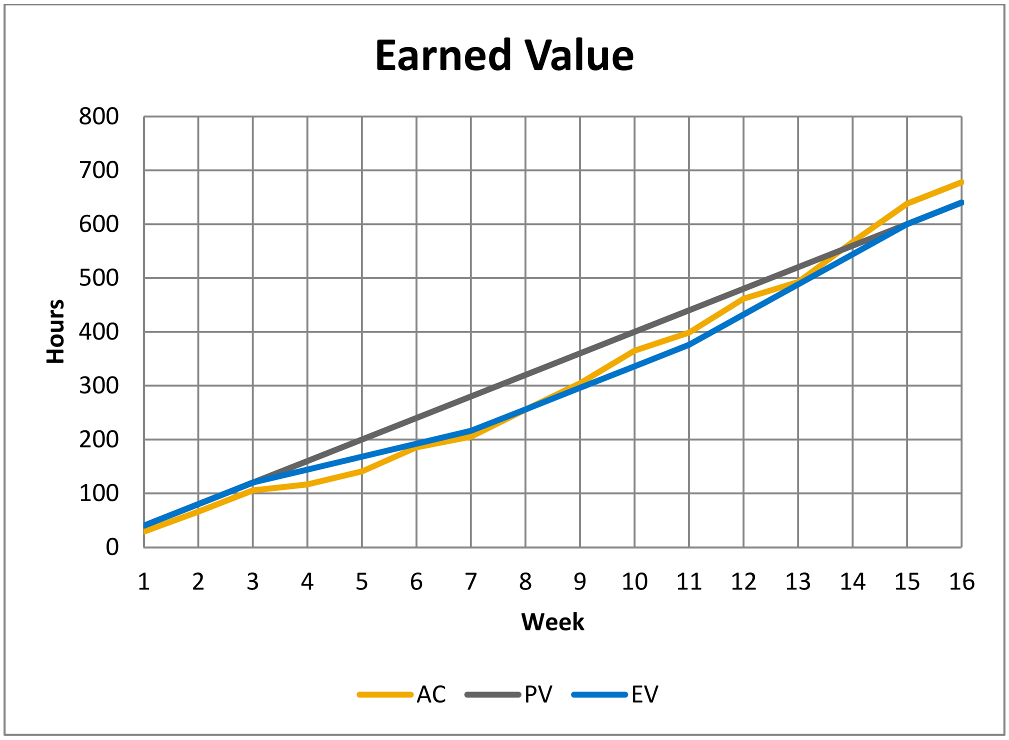


Figure 8: Earned Value

There were significant delays in the project which resulted in the team falling behind schedule. This is shown in Figure 9 below. As it can be seen, the Schedule Variance (SV) line in yellow fell into negative values. The delay in schedule was accompanied by running under cost for the first half of the semester. With no access to data the team had difficulty finding meaningful work on the project. The team received access to the data in week 8 and began putting in extra hours to make up the time. After a reduction in the scope of the project, in part due to data quality, the team was able to deliver the project on time.

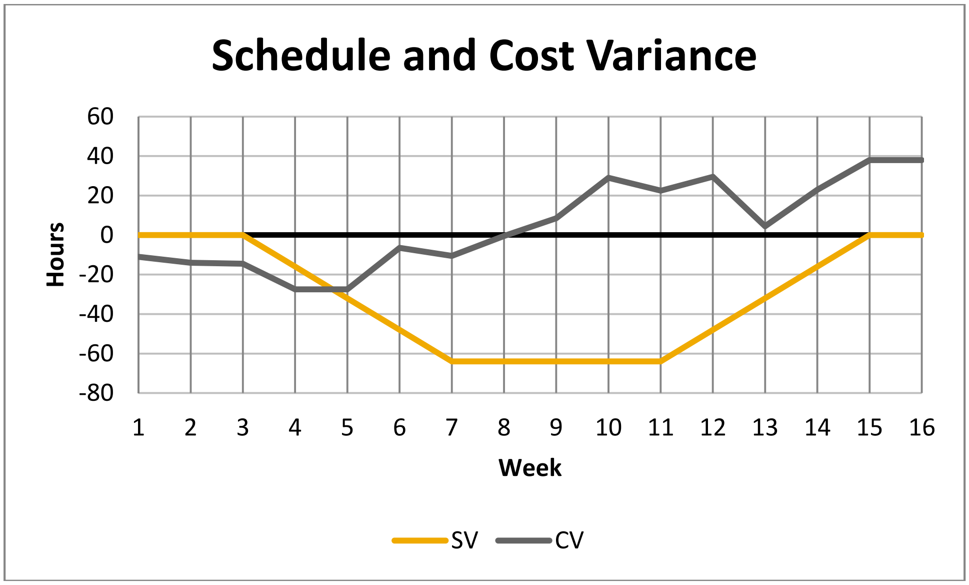


Figure 9: Schedule and Cost Variance

6.3 Risk Assessment and Management

Risk 1: Data Delivery
Description
Due to the data being shipped instead of a digital transfer, there was risk that the data would arrive too late in the semester for the team to develop a fully functional model. If the team did not have access to the data by February 14th, the scope of the model had to be adjusted.

Outcome
The data was not available for analysis until March 7th. As a result the scope of the project was reduced to look for patterns in the mobile web browsing by gender, and develop a model to infer the gender of an unknown subscriber.

Risk 2: Data Access
Description
Depending on the installation of the data there was a chance that the data would only be accessible at the SAP Reston office. The team would need to be escorted while in the office and due to schedule limitations of the project team there would be limited availability to work on data analysis and model.

Outcome
The team was unable to connect to the server with the database from outside of the SAP office. Arturo had building access, because he is interning with SAP, so the team was able to work after hours and on the weekends during the semester.

Risk 3: Big Data Expertise
Description
The team did not have experience with SAP HANA or PAL, or much exposure to data mining techniques at the beginning of the semester. The challenge was two fold in learning the tools and the techniques required.

Outcome
In order to mitigate this risk the team is consulting with Professors at the George Mason University on the technical approach and Subject Matter Experts at SAP on the tools being used.

7 APPENDIX B: WORK BREAKDOWN STRUCTURE

Outline Number	Task Name	Start	Finish	Duration
1	**Project Management**	**23-Jan**	**25-Mar**	**44 days**
1.1	Project Kickoff	23-Jan	23-Jan	1 day
1.2	Create Preliminary Project Description Presentation	23-Jan	27-Jan	3 days
1.3	Preliminary Project Description Presentation	28-Jan	28-Jan	0.38 days
1.4	Meet with Client	31-Jan	31-Jan	0.25 days
1.5	Create Problem Definition and Scope Presentation	23-Jan	27-Jan	3 days
1.6	Problem Definition and Scope Presentation	4-Feb	4-Feb	0 days
1.7	Meet with Dominiconi	7-Feb	7-Feb	0.38 days
1.8	Draft Project Proposal	4-Feb	6-Feb	3 days
1.9	Project Proposal	10-Feb	10-Feb	1 day
1.1	Meeting with SAP PAL team	24-Feb	24-Feb	0.38 days
1.11	Team Meeting with Pablo	27-Feb	27-Feb	0.38 days
1.12	Create Progress Report 1	25-Feb	3-Mar	5 days
1.13	Progress Report 1	4-Mar	4-Mar	1 day
1.14	Create Progress Presentation 2	19-Mar	25-Mar	5 days
1.15	Progress Presentation 2	25-Mar	25-Mar	1 day
1.16	Spring Break	10-Mar	14-Mar	5 days
2	**Research**	**23-Jan**	**28-Feb**	**27 days**
2.1	Mobile Phone Use Demographics	23-Jan	14-Feb	17 days
2.2	Big data Tools	23-Jan	28-Feb	27 days
3	**Model Development**	**23-Jan**	**2-May**	**72 days**
3.1	Get Access To SAP	23-Jan	28-Feb	27 days
3.2	Get Data	23-Jan	6-Mar	31 days
3.3	Determine Approach	7-Mar	17-Mar	7 days

3.4	Analyze Data	7-Mar	22-Apr	33 days
3.5	Develop Model	31-Mar	2-May	25 days
3.6	Sensitivity Analysis	28-Apr	2-May	5 days
4	**Website**	**28-Apr**	**5-May**	**6 days**
4.1	Create Website	28-Apr	2-May	5 days
4.2	Website Due	5-May	5-May	1 day
5	**Final Report**	**16-Apr**	**5-May**	**14 days**
5.1	Draft	16-Apr	29-Apr	10 days
5.2	Review	30-Apr	1-May	2 days
5.3	Tech Edit	2-May	2-May	1 day
5.4	Final Report Due	5-May	5-May	1 day
6	Final Presentation	1-Apr	9-May	29 days
6.1	Draft 1	1-Apr	7-Apr	5 days
6.2	Meet with Professor	8-Apr	8-Apr	1 day
6.3	Meet with Professor	15-Apr	15-Apr	1 day
6.4	Draft 2	16-Apr	17-Apr	2 days

8 APPENDIX C: SQL SCRIPTS AND ALGORITHM GENERATION

Appendix C contains all the SQL scripts that were used to for the team’s ETL and analysis activities. These SQL scripts follow the convention used on the SAP HANA Predictive Analytics Library platform. The team is also providing a summary of the process followed within HANA and Data Services

1. Get female SUB IDs
2. Get male SUB IDs
3. Join lists of males and females into one list
4. Pull all their data from the main table
5. Categorize male and female model training
6. Create View: Five minute durations of categories
7. Usage each day
8. Average daily usage
9. Create table
10. Create set of used female IDs
11. Get 500 female IDs
12. Get 500 male IDs
13. Join lists of males and females into one list
14. Pull all male and female data from the main table
15. Join data with categories
16. Summarize data into 5 minute intervals
17. Average usage each day
18. Average usage for any day
19. Create average usage for any day table
20. Data Services Pivoting
21. Getting training and testing table ready for algorithm
22. PAL algorithm visuals for Bayes
23. PAL algorithm visuals for CHAID24. Create test set mapping
25. Create results
26. Analyze results

1. Get female Subscriber (SUB) IDs

```
CREATE COLUMN TABLE "SCIPAL"."FEMALE_TRAINING" AS (
SELECT top 10000 DISTINCT SUB_ID
FROM "SCI"."sci.db.tables::USAGE_DETAIL_SOURCE"
WHERE DOMAIN IS NOT NULL AND
AGE_BAND <>’Unknown’ AND
AGE_BAND is not null AND
GENDER = ‘F’
)
```

2. Get male SUB IDs

```
CREATE COLUMN TABLE "SCIPAL"."MALE_TESTING" AS (
SELECT top 10000 DISTINCT SUB_ID
FROM "SCI"."sci.db.tables::USAGE_DETAIL_SOURCE"
WHERE DOMAIN IS NOT NULL AND
AGE_BAND <>’Unknown’ AND
AGE_BAND is not null AND
GENDER = ‘M’
)
```

3. Join lists of males and females into one list

```
CREATE COLUMN TABLE "SCIPAL"."BOTH_TRAINING"
AS ((
SELECT *
FROM "SCIPAL"."FEMALE_TRAINING"
) UNION ALL (
SELECT *
FROM "SCIPAL"."MALE_TRAINING"
)
)
```

4. Pull all their data from the main table

```
CREATE COLUMN TABLE "SCIPAL"."BOTH_MODEL_TRAINING" AS (
SELECT a.*
FROM "SCI"."sci.db.tables::USAGE_DETAIL_SOURCE" a
      INNER JOIN "SCIPAL"."BOTH_TRAINING" b
         ON a.SUB_ID = b.SUB_ID)
```

5. Categorize male and female model training

```
CREATE COLUMN TABLE "SCIPAL"."CATEGORIZED_BOTH_MODEL_TRAINING"
AS (
select A.*, B.CATEGORY_1 AS CATEGORY
FROM "SCIPAL"."BOTH_MODEL_TRAINING" A
LEFT JOIN "SCI"."SOURCE_ALL_DISTINCT_DOMAINS_CATEGORIZED" B
ON (a.domain = b.domain)
)
```

6. Create View: Five minute durations of categories

```
CREATE VIEW "SCIPAL"."CATEGORY_DURATION_TRAINING" ( "SUB_ID",
          "GENDER",
         "HANDSET",
         "HOME_POST_CODE",
         "CATEGORY",
          "DAY_OCCURED",
          "RECORD_ID_COUNT",
          "START_TIME_IN_SECONDS",
          "BYTES_IN",
          "BYTES_OUT"
           ) AS
```

```
SELECT SUB_ID, GENDER, HANDSET, HOME_POST_CODE, CATEGORY, EXTRACT(DAY FROM
START_TIME) "DAY",  COUNT(RECORD_ID) "RECORD_ID_COUNT", ROUND(SECONDS_BETWEEN('2014-
01-21', START_TIME)/300,0) "START_TIME_IN_SECONDS", SUM(BYTES_IN) "SUM_BYTES_IN",
SUM(BYTES_OUT) "SUM_BYTES_OUT"
FROM "SCIPAL"."CATEGORIZED_BOTH_MODEL_TRAINING"
WHERE CATEGORY IS NOT NULL
GROUP BY SUB_ID, HANDSET, HOME_POST_CODE, GENDER, CATEGORY, EXTRACT(DAY FROM
START_TIME), ROUND(SECONDS_BETWEEN('2014-01-21', START_TIME)/300,0)
WITH READ ONLY
```

7. Usage each day

```
CREATE VIEW "SCIPAL"."CATEGORY_DURATION_PER_DAY_TRAINING" AS (
SELECT SUB_ID, GENDER, HANDSET, HOME_POST_CODE, CATEGORY, DAY_OCCURED,
SUM(RECORD_ID_COUNT) "WEIGHT" ,
COUNT(RECORD_ID_COUNT)*5 "DURATION_SPAN",
SUM(BYTES_IN) "BYTES_IN" ,
SUM(BYTES_OUT) "BYTES_OUT"
FROM (
--"SCIPAL"."FIVE_MIN_DURATIONS_OF_CATEGORIES"
SELECT SUB_ID, GENDER, HANDSET, HOME_POST_CODE, CATEGORY, EXTRACT(DAY FROM
START_TIME) "DAY_OCCURED",  COUNT(RECORD_ID) "RECORD_ID_COUNT",
ROUND(SECONDS_BETWEEN('2014-01-21', START_TIME)/300,0) "START_TIME_IN_SECONDS",
SUM(BYTES_IN) "BYTES_IN", SUM(BYTES_OUT) "BYTES_OUT"
FROM "SCIPAL"."CATEGORIZED_BOTH_MODEL_TRAINING"
WHERE CATEGORY IS NOT NULL
GROUP BY SUB_ID, HANDSET, HOME_POST_CODE, GENDER, CATEGORY, EXTRACT(DAY FROM
START_TIME), ROUND(SECONDS_BETWEEN('2014-01-21', START_TIME)/300,0)
)
GROUP BY SUB_ID, GENDER, HANDSET, HOME_POST_CODE, CATEGORY, DAY_OCCURED
)
```

8. Average daily usage

```
CREATE VIEW "SCIPAL"."CATEGORY_DURATION_AVG_TRAINING" AS (
SELECT SUB_ID, GENDER, HANDSET, HOME_POST_CODE, CATEGORY,
AVG(WEIGHT) "WEIGHT" ,
AVG(DURATION_SPAN) "DURATION_SPAN",
AVG(BYTES_IN) "BYTES_IN" ,
AVG(BYTES_OUT) "BYTES_OUT"
FROM "SCIPAL"."CATEGORY_DURATION_PER_DAY_TRAINING"
GROUP BY SUB_ID, GENDER, HANDSET, HOME_POST_CODE, CATEGORY
)
```

9. Create table

SUB_ID, CATEGORY, WEIGHT, DURATION_SPAN, HANDSET, HOME_POST_CODE, GENDER

```
CREATE COLUMN TABLE "SCIPAL"."PRETRANS_CATEGORIZED_BOTH_MODEL_TRAINING_2" AS
(SELECT *
FROM "SCIPAL"."CATEGORY_DURATION_AVG_TRAINING")
```

10. Create set of used female IDs

```
CREATE COLUMN TABLE "SCIPAL"."FEMALE_USED_TRAINING_SUBIDS" AS (
(
SELECT *
FROM "SCIPAL"."FEMALE_TRAINING"
) UNION ALL (
SELECT *
FROM "SCIPAL"."FEMALE_TRAINING_2"
)
)
```

11. Get 500 female IDs

```
CREATE COLUMN TABLE "SCIPAL"."FEMALE_TESTING" AS (

SELECT top 500 DISTINCT a.SUB_ID
FROM "SCI"."sci.db.tables::USAGE_DETAIL_SOURCE" a
LEFT JOIN "SCIPAL"."FEMALE_USED_TRAINING_SUBIDS" b
ON a.SUB_ID = b.SUB_ID
WHERE b.SUB_ID IS NULL AND
a.DOMAIN IS NOT NULL AND
a.AGE_BAND <>’Unknown’ AND
a.AGE_BAND is not null AND
a.GENDER = ‘F’
)
```

12. Get 500 male IDs

```
CREATE COLUMN TABLE "SCIPAL"."MALE_TESTING" AS (

SELECT top 500 DISTINCT a.SUB_ID
FROM "SCI"."sci.db.tables::USAGE_DETAIL_SOURCE" a
LEFT JOIN "SCIPAL"."MALE_TRAINING" b
ON a.SUB_ID = b.SUB_ID
WHERE b.SUB_ID IS NULL AND
a.DOMAIN IS NOT NULL AND
a.AGE_BAND <>’Unknown’ AND
a.AGE_BAND is not null AND
a.GENDER = ‘M’
)
```

13. Join lists of males and females into one list

```
CREATE COLUMN TABLE "SCIPAL"."BOTH_TESTING"
AS ((
SELECT *
FROM "SCIPAL"."FEMALE_TESTING"
) UNION ALL (
SELECT *
FROM "SCIPAL"."MALE_TESTING"
)
)
```

14. Pull all male and female data from the main table

```
CREATE COLUMN TABLE "SCIPAL"."BOTH_MODEL_TESTING" AS (
SELECT a.*
FROM "SCI"."sci.db.tables::USAGE_DETAIL_SOURCE" a
    INNER JOIN "SCIPAL"."BOTH_TESTING" b
      ON a.SUB_ID = b.SUB_ID)
```

15. Join data with categories

```
CREATE COLUMN TABLE "SCIPAL"."CATEGORIZED_BOTH_MODEL_TESTING"
AS (
select A.*, B.CATEGORY_1 AS CATEGORY
FROM "SCIPAL"."BOTH_MODEL_TESTING" A
LEFT JOIN "SCI"."SOURCE_ALL_DISTINCT_DOMAINS_CATEGORIZED" B
ON (A.DOMAIN = B.DOMAIN)
)
```

16. Summarize data into 5 minute intervals

```
CREATE VIEW "SCIPAL"."CATEGORY_DURATION_TESTING" ( "SUB_ID",
        "GENDER",

       "HANDSET",
       "HOME_POST_CODE",
       "CATEGORY",
        "DAY_OCCURED",
        "RECORD_ID_COUNT",
        "START_TIME_IN_SECONDS",
        "BYTES_IN",
        "BYTES_OUT"
         ) AS
SELECT SUB_ID, GENDER, HANDSET, HOME_POST_CODE, CATEGORY, EXTRACT(DAY FROM
START_TIME) "DAY",  COUNT(RECORD_ID) "RECORD_ID_COUNT", ROUND(SECONDS_BETWEEN('2014-
01-21', START_TIME)/300,0) "START_TIME_IN_SECONDS", SUM(BYTES_IN) "SUM_BYTES_IN",
SUM(BYTES_OUT) "SUM_BYTES_OUT"
FROM "SCIPAL"."CATEGORIZED_BOTH_MODEL_TESTING"
WHERE CATEGORY IS NOT NULL
GROUP BY SUB_ID, HANDSET, HOME_POST_CODE, GENDER, CATEGORY, EXTRACT(DAY FROM
START_TIME), ROUND(SECONDS_BETWEEN('2014-01-21', START_TIME)/300,0)
WITH READ ONLY
```

17. Average usage each day

```
CREATE VIEW "SCIPAL"."CATEGORY_DURATION_PER_DAY_TESTING" AS (
SELECT SUB_ID, GENDER, HANDSET, HOME_POST_CODE, CATEGORY, DAY_OCCURED,

SUM(RECORD_ID_COUNT) "WEIGHT" ,
COUNT(RECORD_ID_COUNT)*5 "DURATION_SPAN",
SUM(BYTES_IN) "BYTES_IN" ,
SUM(BYTES_OUT) "BYTES_OUT"
FROM (
--"SCIPAL"."FIVE_MIN_DURATIONS_OF_CATEGORIES_TESTING"
SELECT SUB_ID, GENDER, HANDSET, HOME_POST_CODE, CATEGORY, EXTRACT(DAY FROM
START_TIME) "DAY_OCCURED",  COUNT(RECORD_ID) "RECORD_ID_COUNT",
ROUND(SECONDS_BETWEEN('2014-01-21', START_TIME)/300,0) "START_TIME_IN_SECONDS",
SUM(BYTES_IN) "BYTES_IN", SUM(BYTES_OUT) "BYTES_OUT"
FROM "SCIPAL"."CATEGORIZED_BOTH_MODEL_TESTING"
WHERE CATEGORY IS NOT NULL
GROUP BY SUB_ID, HANDSET, HOME_POST_CODE, GENDER, CATEGORY, EXTRACT(DAY FROM
START_TIME), ROUND(SECONDS_BETWEEN('2014-01-21', START_TIME)/300,0)
)

GROUP BY SUB_ID, GENDER, HANDSET, HOME_POST_CODE, CATEGORY, DAY_OCCURED
)
```

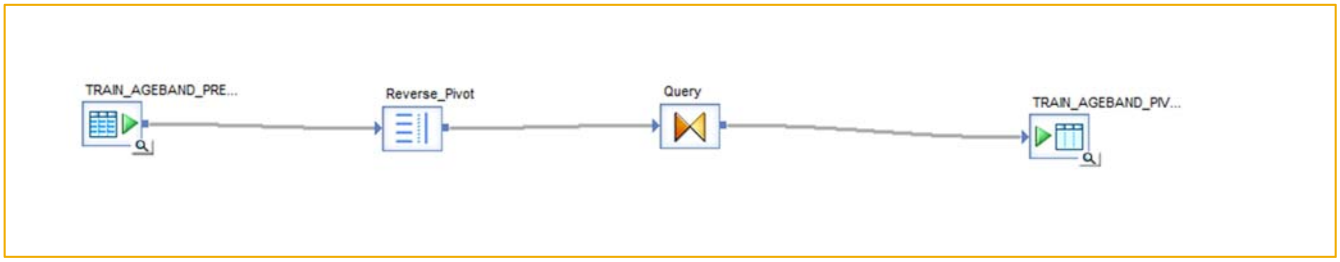
18. Average usage for any day

```
CREATE VIEW "SCIPAL"."CATEGORY_DURATION_AVG_TESTING" AS (
SELECT SUB_ID, GENDER, HANDSET, HOME_POST_CODE, CATEGORY,
AVG(WEIGHT) "WEIGHT" ,
AVG(DURATION_SPAN) "DURATION_SPAN",
AVG(BYTES_IN) "BYTES_IN" ,
AVG(BYTES_OUT) "BYTES_OUT"
FROM "SCIPAL"."CATEGORY_DURATION_PER_DAY_TESTING"
GROUP BY SUB_ID, GENDER, HANDSET, HOME_POST_CODE, CATEGORY
)
```

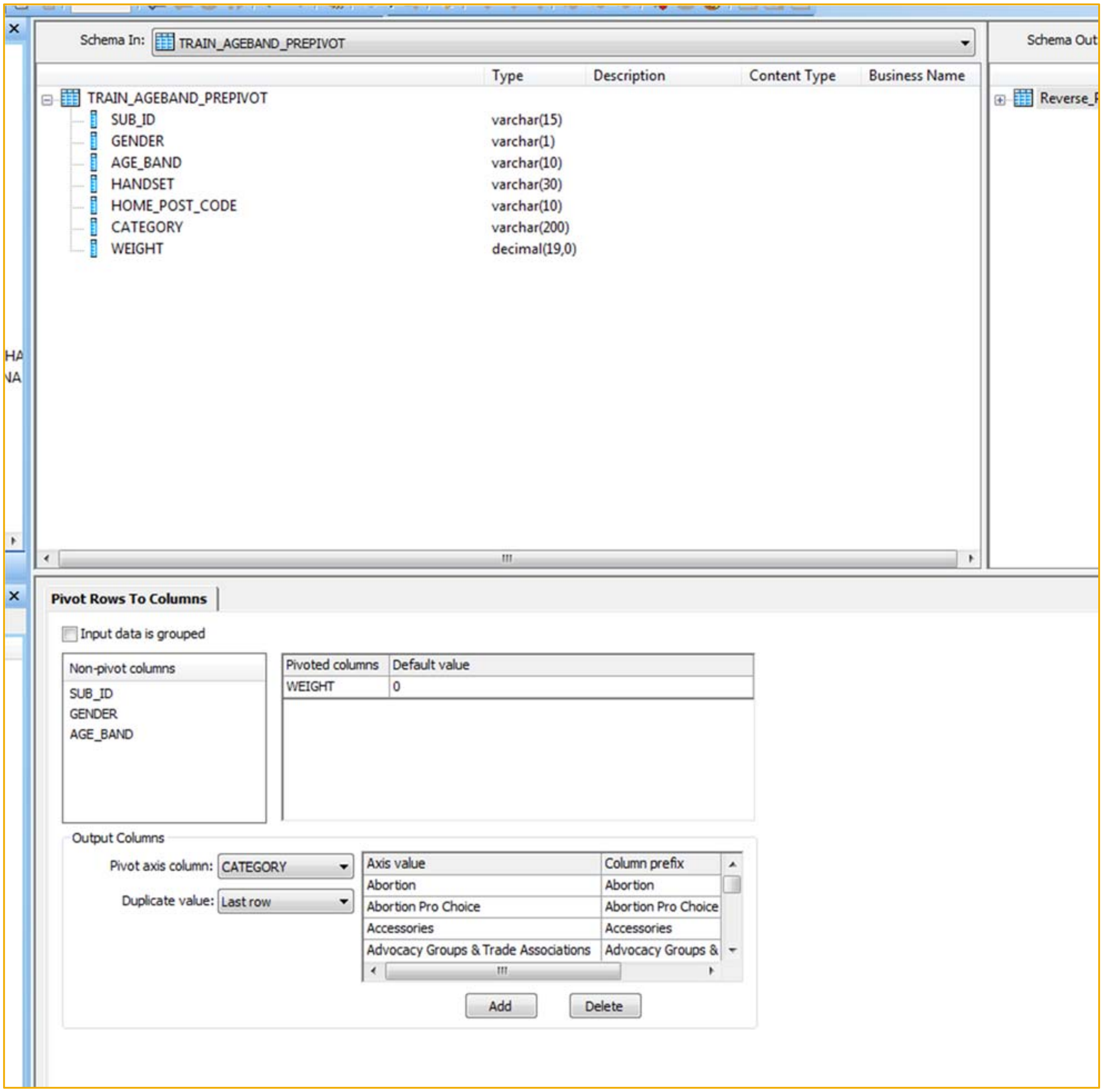
19. Create average usage for any day table

```
CREATE COLUMN TABLE "SCIPAL"."PRETRANS_CATEGORIZED_BOTH_MODEL_TESTING" AS (SELECT
*
FROM "SCIPAL"."CATEGORY_DURATION_AVG_TESTING")
```

20. Data Services Pivoting



Pivot Structure



Schema In: TRAIN_AGEBAND_PREPIVOT

Schema Out

	Type	Description	Content Type	Business Name
TRAIN_AGEBAND_PREPIVOT				
SUB_ID	varchar(15)			
GENDER	varchar(1)			
AGE_BAND	varchar(10)			
HANDSET	varchar(30)			
HOME_POST_CODE	varchar(10)			
CATEGORY	varchar(200)			
WEIGHT	decimal(19,0)			

Reverse_P

Pivot Rows To Columns

☐ Input data is grouped

Non-pivot columns
SUB_ID
GENDER
AGE_BAND

Pivoted columns	Default value
WEIGHT	0

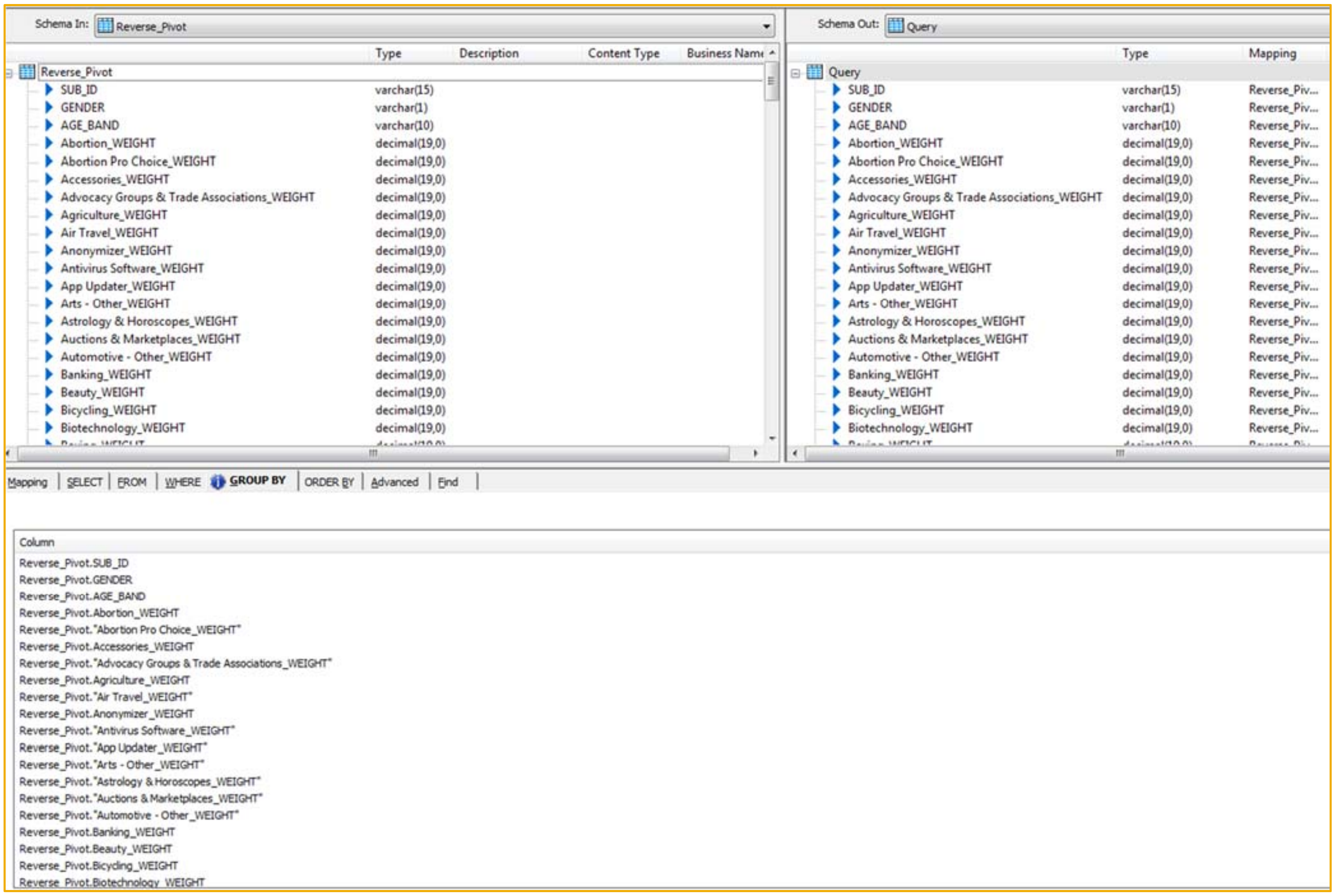
Output Columns

Pivot axis column: CATEGORY

Duplicate value: Last row

Axis value	Column prefix
Abortion	Abortion
Abortion Pro Choice	Abortion Pro Choice
Accessories	Accessories
Advocacy Groups & Trade Associations	Advocacy Groups &

Reverse Plot and Manual Category Inputs



Schema In: Reverse_Pivot

	Type	Description	Content Type	Business Name
Reverse_Pivot				
SUB_ID	varchar(15)			
GENDER	varchar(1)			
AGE_BAND	varchar(10)			
Abortion_WEIGHT	decimal(19,0)			
Abortion Pro Choice_WEIGHT	decimal(19,0)			
Accessories_WEIGHT	decimal(19,0)			
Advocacy Groups & Trade Associations_WEIGHT	decimal(19,0)			
Agriculture_WEIGHT	decimal(19,0)			
Air Travel_WEIGHT	decimal(19,0)			
Anonymizer_WEIGHT	decimal(19,0)			
Antivirus Software_WEIGHT	decimal(19,0)			
App Updater_WEIGHT	decimal(19,0)			
Arts - Other_WEIGHT	decimal(19,0)			
Astrology & Horoscopes_WEIGHT	decimal(19,0)			
Auctions & Marketplaces_WEIGHT	decimal(19,0)			
Automotive - Other_WEIGHT	decimal(19,0)			
Banking_WEIGHT	decimal(19,0)			
Beauty_WEIGHT	decimal(19,0)			
Bicycling_WEIGHT	decimal(19,0)			
Biotechnology_WEIGHT	decimal(19,0)			

Schema Out: Query

	Type	Mapping
Query		
SUB_ID	varchar(15)	Reverse_Piv...
GENDER	varchar(1)	Reverse_Piv...
AGE_BAND	varchar(10)	Reverse_Piv...
Abortion_WEIGHT	decimal(19,0)	Reverse_Piv...
Abortion Pro Choice_WEIGHT	decimal(19,0)	Reverse_Piv...
Accessories_WEIGHT	decimal(19,0)	Reverse_Piv...
Advocacy Groups & Trade Associations_WEIGHT	decimal(19,0)	Reverse_Piv...
Agriculture_WEIGHT	decimal(19,0)	Reverse_Piv...
Air Travel_WEIGHT	decimal(19,0)	Reverse_Piv...
Anonymizer_WEIGHT	decimal(19,0)	Reverse_Piv...
Antivirus Software_WEIGHT	decimal(19,0)	Reverse_Piv...
App Updater_WEIGHT	decimal(19,0)	Reverse_Piv...
Arts - Other_WEIGHT	decimal(19,0)	Reverse_Piv...
Astrology & Horoscopes_WEIGHT	decimal(19,0)	Reverse_Piv...
Auctions & Marketplaces_WEIGHT	decimal(19,0)	Reverse_Piv...
Automotive - Other_WEIGHT	decimal(19,0)	Reverse_Piv...
Banking_WEIGHT	decimal(19,0)	Reverse_Piv...
Beauty_WEIGHT	decimal(19,0)	Reverse_Piv...
Bicycling_WEIGHT	decimal(19,0)	Reverse_Piv...
Biotechnology_WEIGHT	decimal(19,0)	Reverse_Piv...

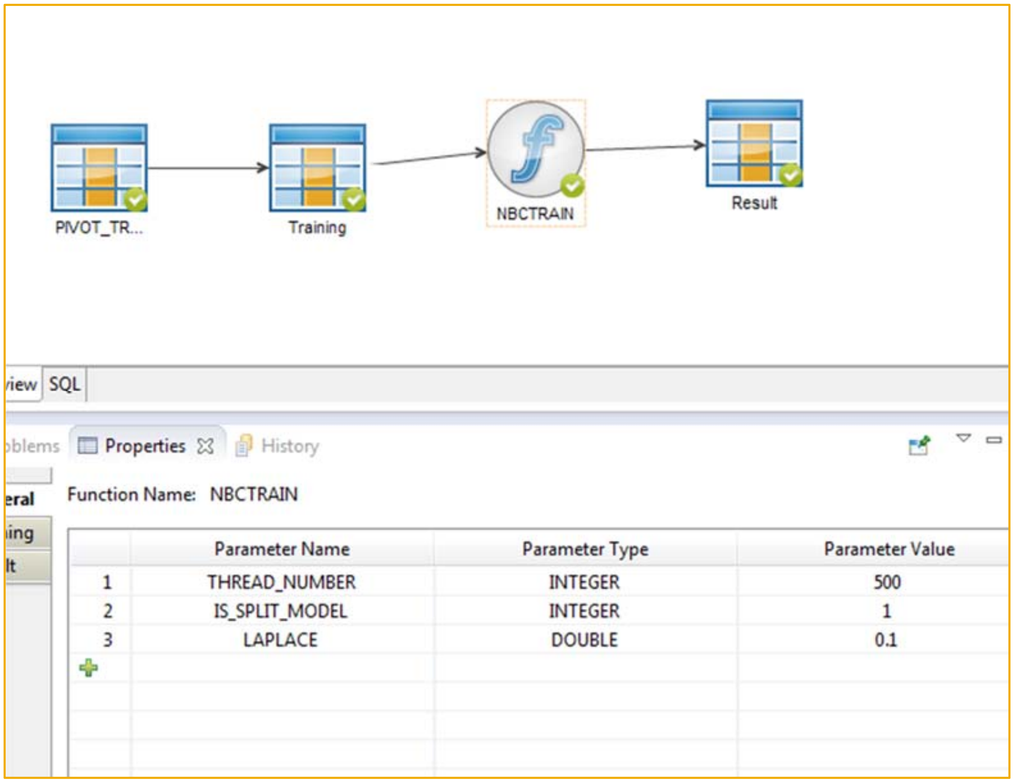
Column
Reverse_Pivot.SUB_ID
Reverse_Pivot.GENDER
Reverse_Pivot.AGE_BAND
Reverse_Pivot.Abortion_WEIGHT
Reverse_Pivot."Abortion Pro Choice_WEIGHT"
Reverse_Pivot.Accessories_WEIGHT
Reverse_Pivot."Advocacy Groups & Trade Associations_WEIGHT"
Reverse_Pivot.Agriculture_WEIGHT
Reverse_Pivot."Air Travel_WEIGHT"
Reverse_Pivot.Anonymizer_WEIGHT
Reverse_Pivot."Antivirus Software_WEIGHT"
Reverse_Pivot."App Updater_WEIGHT"
Reverse_Pivot."Arts - Other_WEIGHT"
Reverse_Pivot."Astrology & Horoscopes_WEIGHT"
Reverse_Pivot."Auctions & Marketplaces_WEIGHT"
Reverse_Pivot."Automotive - Other_WEIGHT"
Reverse_Pivot.Banking_WEIGHT
Reverse_Pivot.Beauty_WEIGHT
Reverse_Pivot.Bicycling_WEIGHT
Reverse_Pivot.Biotechnology_WEIGHT

Query Table and Assigning Output Results

21. Getting training and testing table ready for algorithm

```
CREATE COLUMN TABLE "SCIPAL_2"."MASTER_TRAIN_ALG_READY" (
"ABORTION",
"Abortion_Pro_Choice",
"ACCESSORIES",
"Advocacy_Groups_Trade_Associations",
"AGRICULTURE",
"Air_Travel",
"ANONYMIZER",
"Antivirus_Software",
"App_Updater",
"Arts_Other",
"Astrology_Horoscopes",
"Auctions_Marketplaces",
"Automotive_Other",
"BANKING",
"BEAUTY",
"BICYCLING",
"BIOTECHNOLOGY",
"BOXING",
"Business_Other",
"Careers_Other",
          "GENDER")
AS (SELECT
TO_VARCHAR("ABORTION_WEIGHT"),
TO_VARCHAR("Abortion Pro Choice_WEIGHT"),
TO_VARCHAR("ACCESSORIES_WEIGHT"),
TO_VARCHAR("Advocacy Groups & Trade Associations_WEIGHT"),
TO_VARCHAR("AGRICULTURE_WEIGHT"),
TO_VARCHAR("Air Travel_WEIGHT"),
TO_VARCHAR("ANONYMIZER_WEIGHT"),
TO_VARCHAR("Antivirus Software_WEIGHT"),
TO_VARCHAR("App Updater_WEIGHT"),
TO_VARCHAR("Arts - Other_WEIGHT"),
TO_VARCHAR("Astrology & Horoscopes_WEIGHT"),
TO_VARCHAR("Auctions & Marketplaces_WEIGHT"),
TO_VARCHAR("Automotive - Other_WEIGHT"),
TO_VARCHAR("BANKING_WEIGHT"),
TO_VARCHAR("BEAUTY_WEIGHT"),
TO_VARCHAR("BICYCLING_WEIGHT"),
TO_VARCHAR("BIOTECHNOLOGY_WEIGHT"),
TO_VARCHAR("BOXING_WEIGHT"),
TO_VARCHAR("Business - Other_WEIGHT"),
TO_VARCHAR("Careers - Other_WEIGHT"),
          TO_VARCHAR("GENDER")
FROM "SCIPAL_2"."EXAMPLE_TRAIN_PIVOT")
          ------------------------------------------------------------
CREATE COLUMN TABLE "SCIPAL_2"."MASTER_TEST_ALG_READY" (
"ID"
"ABORTION",
"Abortion_Pro_Choice",
"ACCESSORIES",
"Advocacy_Groups_Trade_Associations",
"AGRICULTURE",
"Air_Travel",
"ANONYMIZER",
"Antivirus_Software",
"App_Updater",
"Arts_Other",
"Astrology_Horoscopes",
"Auctions_Marketplaces",
"Automotive_Other",
"BANKING",
"BEAUTY",
"BICYCLING",
"BIOTECHNOLOGY",
"BOXING",
"Business_Other",
"Careers_Other")
AS (SELECT
TO_INTEGER(row_number() OVER (ORDER BY SUB_ID ASC, HANDSET_CLEAN ASC)) as ID,
TO_VARCHAR("ABORTION_WEIGHT"),
TO_VARCHAR("Abortion Pro Choice_WEIGHT"),
TO_VARCHAR("ACCESSORIES_WEIGHT"),
TO_VARCHAR("Advocacy Groups & Trade Associations_WEIGHT"),
TO_VARCHAR("AGRICULTURE_WEIGHT"),
TO_VARCHAR("Air Travel_WEIGHT"),
TO_VARCHAR("ANONYMIZER_WEIGHT"),
TO_VARCHAR("Antivirus Software_WEIGHT"),
TO_VARCHAR("App Updater_WEIGHT"),
TO_VARCHAR("Arts - Other_WEIGHT"),
TO_VARCHAR("Astrology & Horoscopes_WEIGHT"),
TO_VARCHAR("Auctions & Marketplaces_WEIGHT"),
TO_VARCHAR("Automotive - Other_WEIGHT"),
TO_VARCHAR("BANKING_WEIGHT"),
TO_VARCHAR("BEAUTY_WEIGHT"),
TO_VARCHAR("BICYCLING_WEIGHT"),
TO_VARCHAR("BIOTECHNOLOGY_WEIGHT"),
TO_VARCHAR("BOXING_WEIGHT"),
TO_VARCHAR("Business - Other_WEIGHT"),
TO_VARCHAR("Careers - Other_WEIGHT")
FROM "SCIPAL_2"."EXAMPLE_TEST_PIVOT")
```

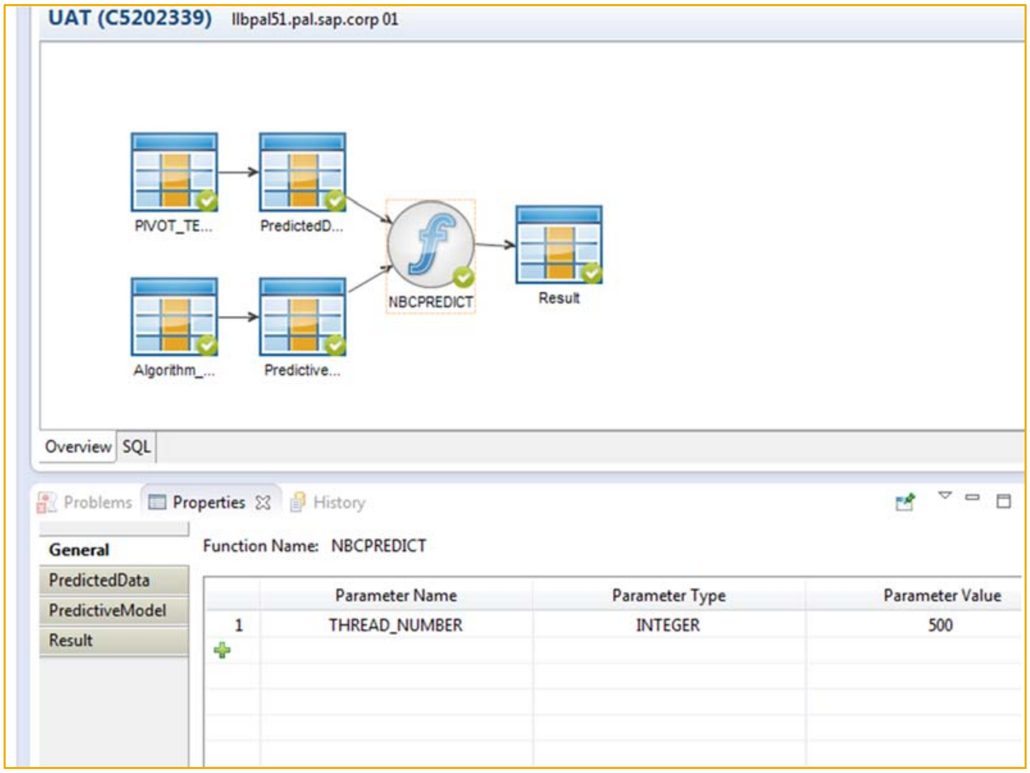
22. PAL algorithm visuals for Bayes



Function Name: NBCTRAIN

	Parameter Name	Parameter Type	Parameter Value
1	THREAD_NUMBER	INTEGER	500
2	IS_SPLIT_MODEL	INTEGER	1
3	LAPLACE	DOUBLE	0.1

Bayes Training Parameters

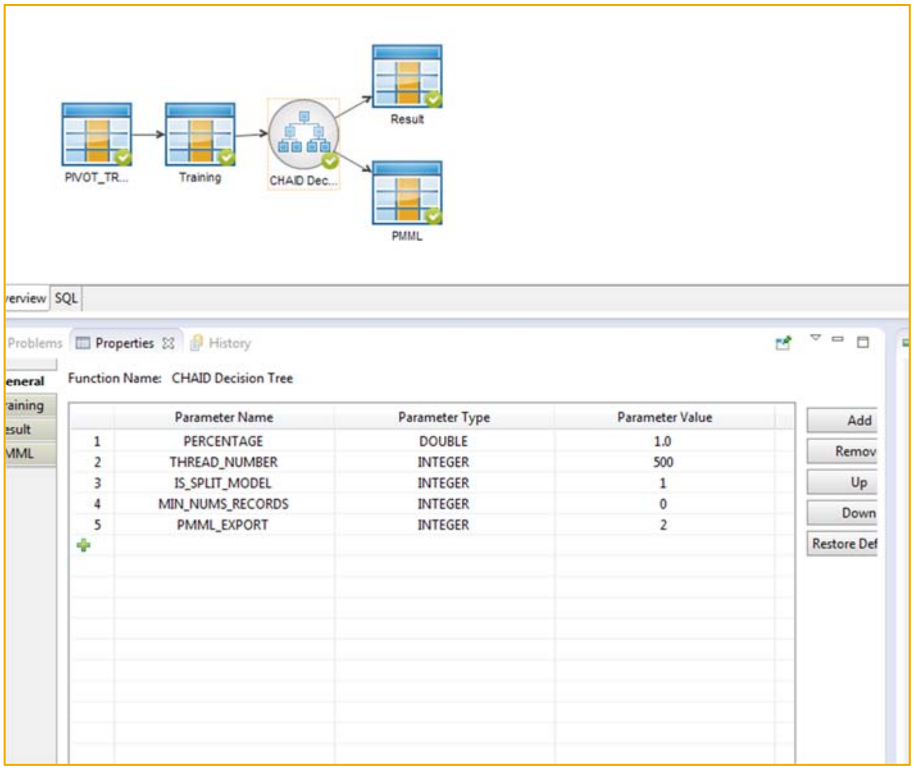


Function Name: NBCPREDICT

	Parameter Name	Parameter Type	Parameter Value
1	THREAD_NUMBER	INTEGER	500

Bayes Testing and Predication Parameters

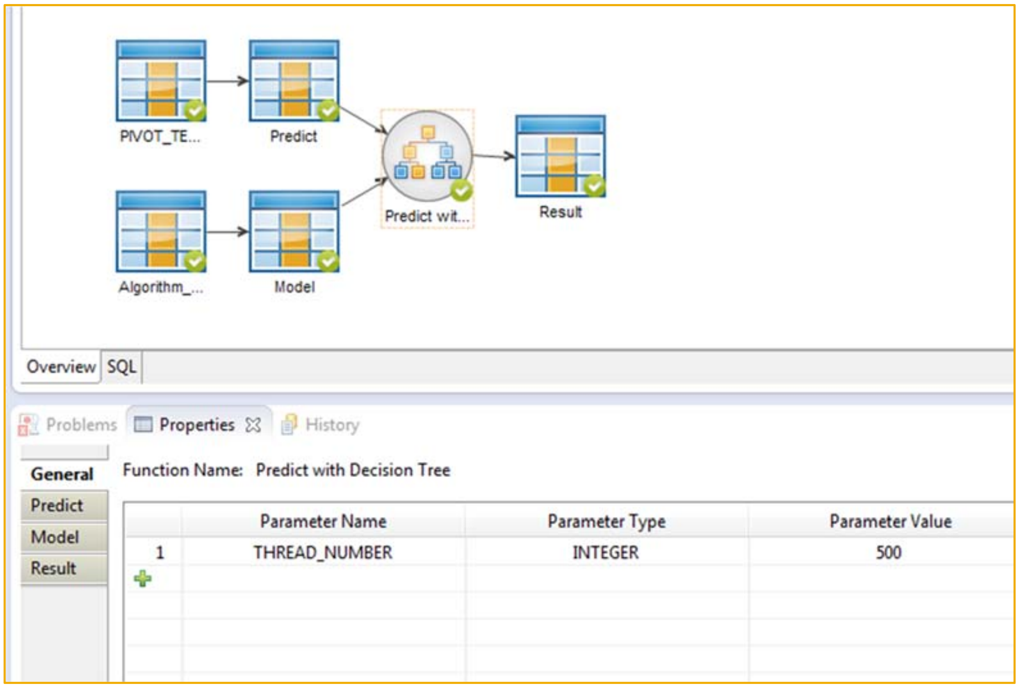
23. PAL algorithm visuals for CHAID



Function Name: CHAID Decision Tree

	Parameter Name	Parameter Type	Parameter Value
1	PERCENTAGE	DOUBLE	1.0
2	THREAD_NUMBER	INTEGER	500
3	IS_SPLIT_MODEL	INTEGER	1
4	MIN_NUMS_RECORDS	INTEGER	0
5	PMML_EXPORT	INTEGER	2

Chaid Training Parameters



Function Name: Predict with Decision Tree

	Parameter Name	Parameter Type	Parameter Value
1	THREAD_NUMBER	INTEGER	500

Chaid Testing and Prediction Parameters

24. Create test set mapping

```
create view "SCIPAL"."MAPPING_C4_SIMPLE"
("ID", "GENDER", "SUB_ID", "HANDSET_CLEAN") AS SELECT
TO_INTEGER(row_number() OVER (
ORDER BY SUB_ID asc, HANDSET_CLEAN asc)) as ID, GENDER, SUB_ID, HANDSET_CLEAN
FROM "SCIPAL"."ZZZ_SIMPLE_TABLE_TESTING"
```

25. Create results

```
CREATE COLUMN TABLE "SCIPAL"."C4_SIMPLE_RESULTS"
AS (
select A.*, B.GENDER AS TEST_GENDER
FROM "SCIPAL"."MAPPING_C4_SIMPLE" A
LEFT JOIN "SCIPAL"."Algorithm_Project::C45_PREDICT.C45_PREDICT_predictWithDT_Result" B
ON (a.ID = b.ID)
)
```

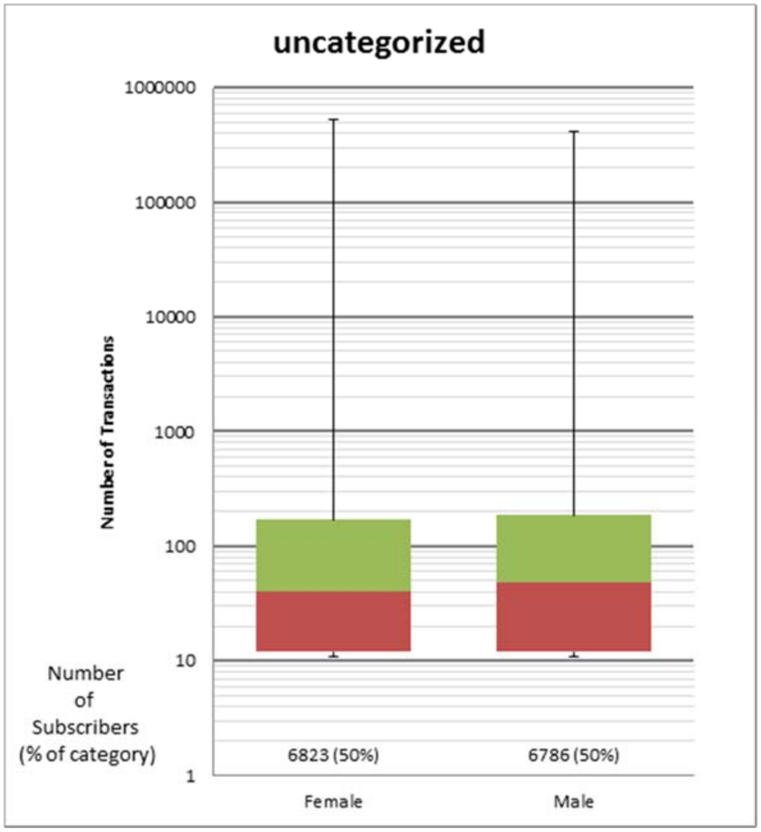
26. Analyze results

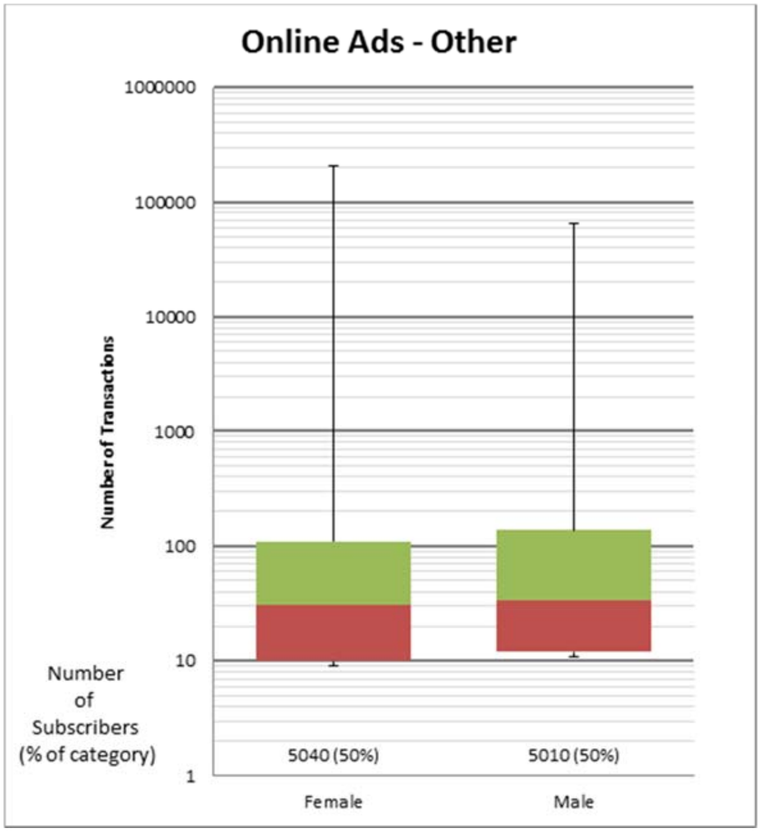
```
SELECT COUNT (test_gender)
FROM "SCIPAL"."C4_SIMPLE_RESULTS"
WHERE gender = 'F' AND test_gender = 'F'
```

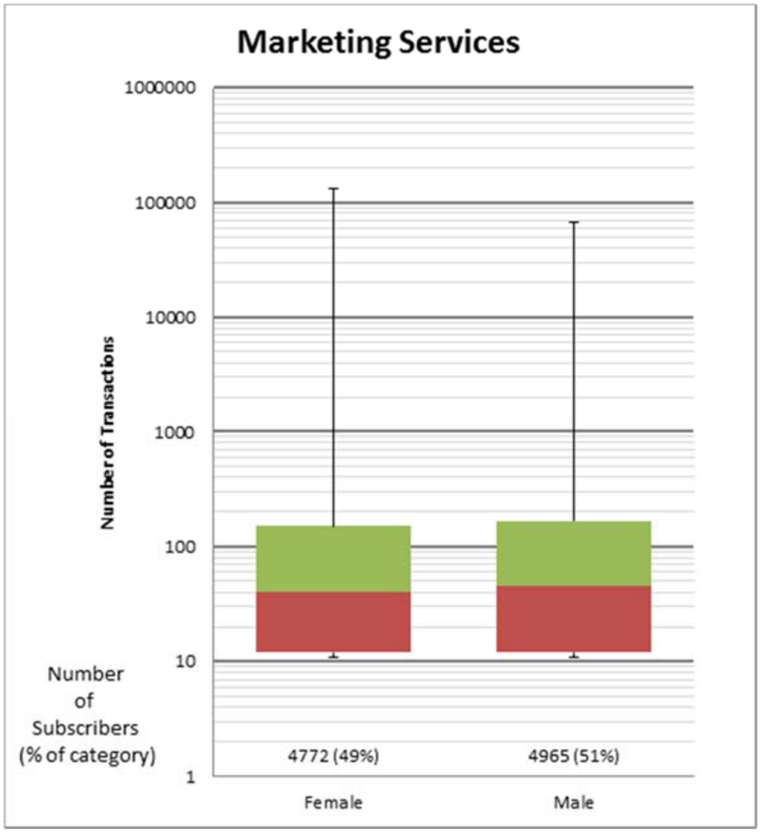
9 APPENDIX D: BOX PLOTS FOR TOP 20 CATEGORIES

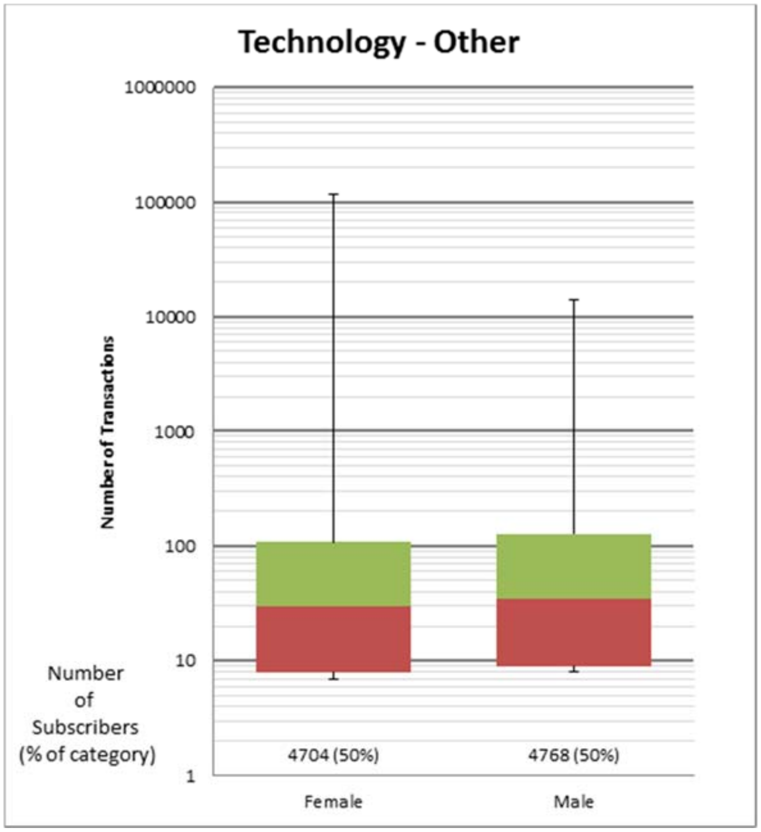
The red and green coloring of the box plots were a function of the graphics software used to plot the data, and do not hold any significance other than differentiating the middle quartiles.

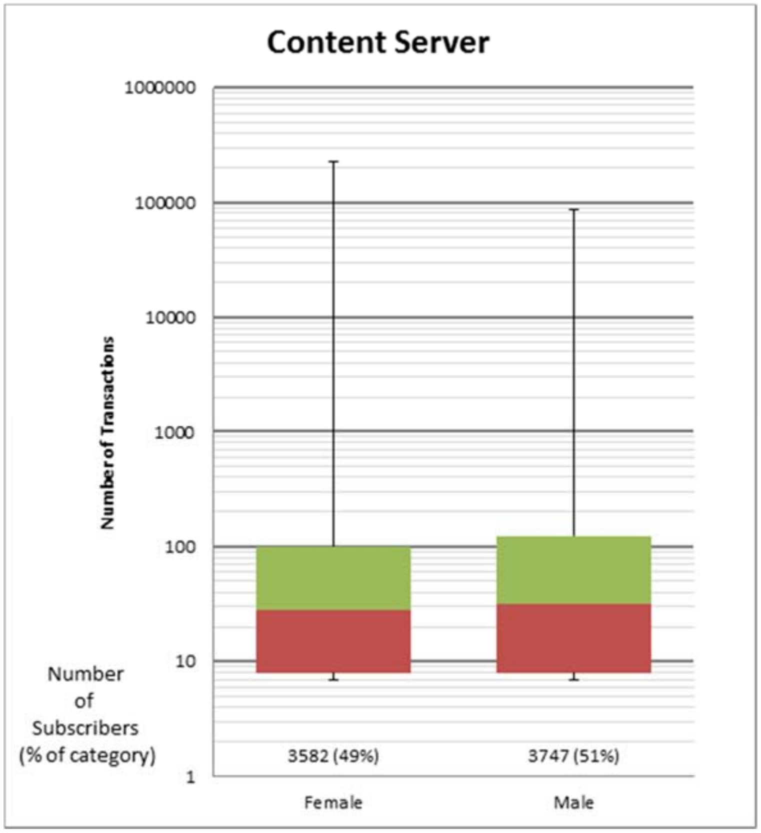
Rank	Category
1.	Uncategorized
2.	Online Ads - Other
3.	Marketing Services
4.	Technology - Other
5.	Content Server
6.	Games
7.	News
8.	Portal Sites
9.	Information Security
10.	File Repositories
11.	Streaming & Downloadable Video
12.	Business - Other
13.	Computer Peripherals
14.	Personal Pages & Blogs
15.	Entertainment - Other
16.	Travel - Other
17.	Community Forums
18.	Social Networking
19.	Mobile Phones
20.	Online Shopping

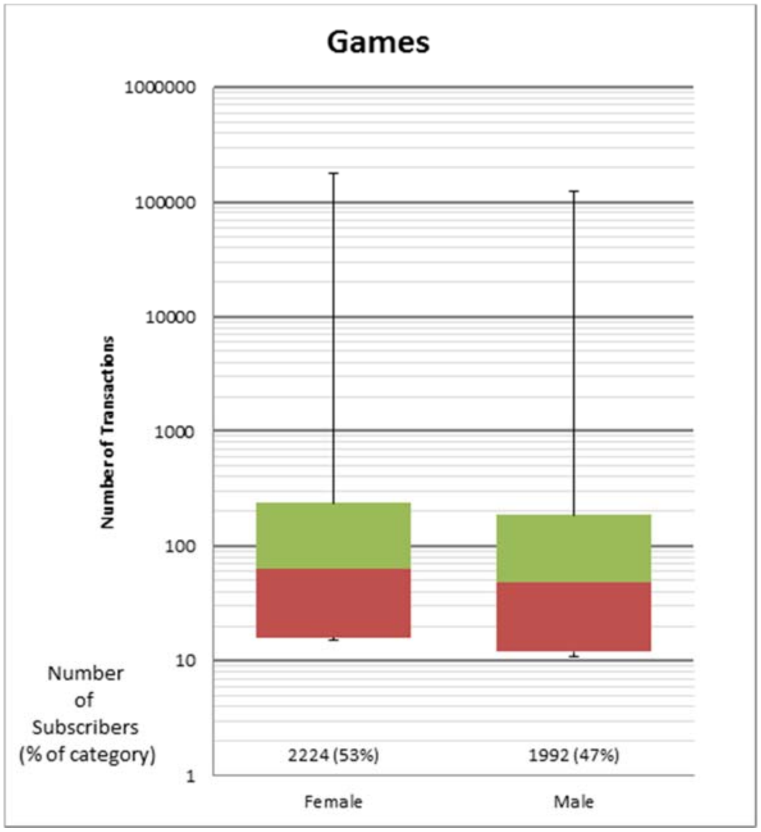

uncategorized
1000000
100000
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
6823 (50%)
6786 (50%)
Female
Male

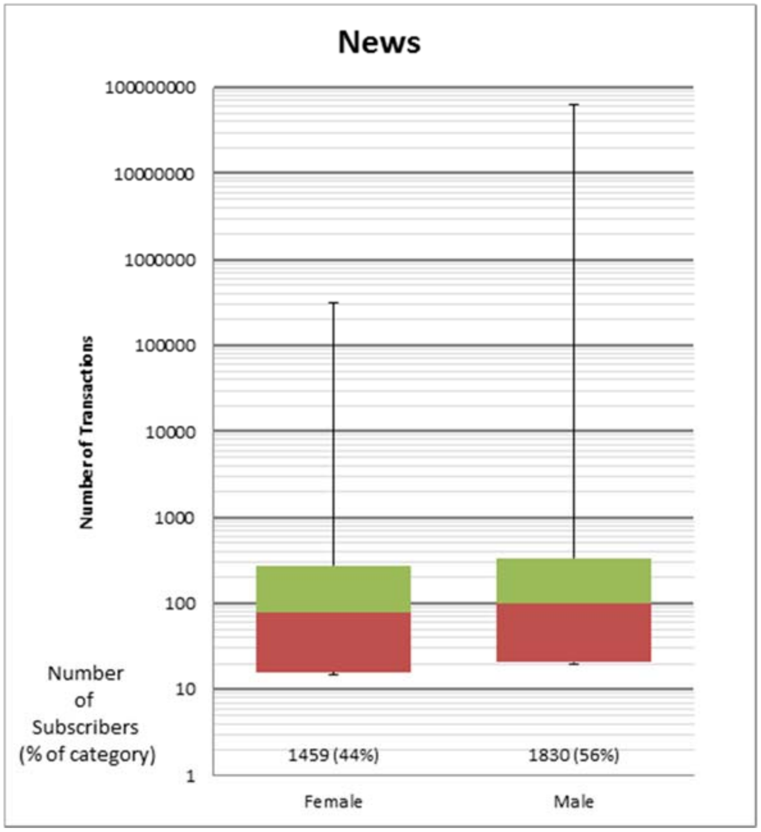

Online Ads - Other
1000000
100000
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
5040 (50%)
5010 (50%)
Female
Male

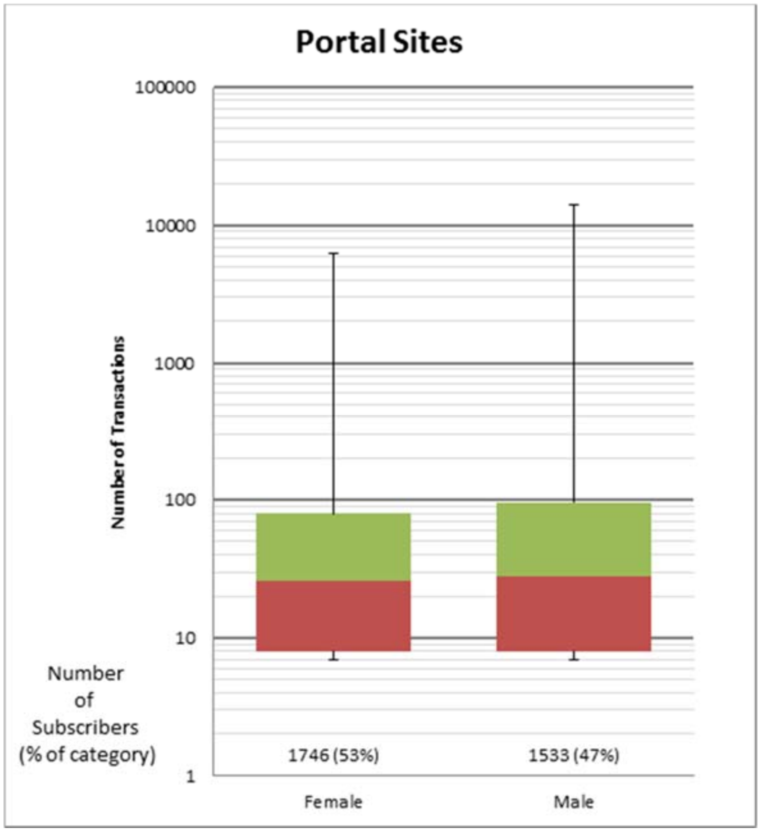

Marketing Services
1000000
100000
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
4772 (49%)
4965 (51%)
Female
Male

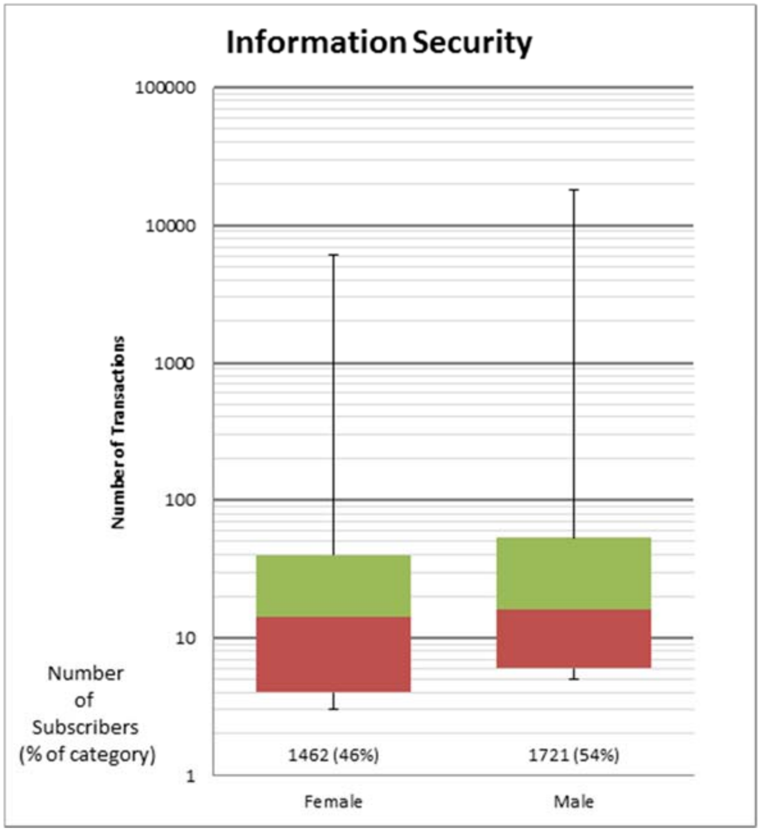

Technology - Other
1000000
100000
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
4704 (50%)
4768 (50%)
Female
Male

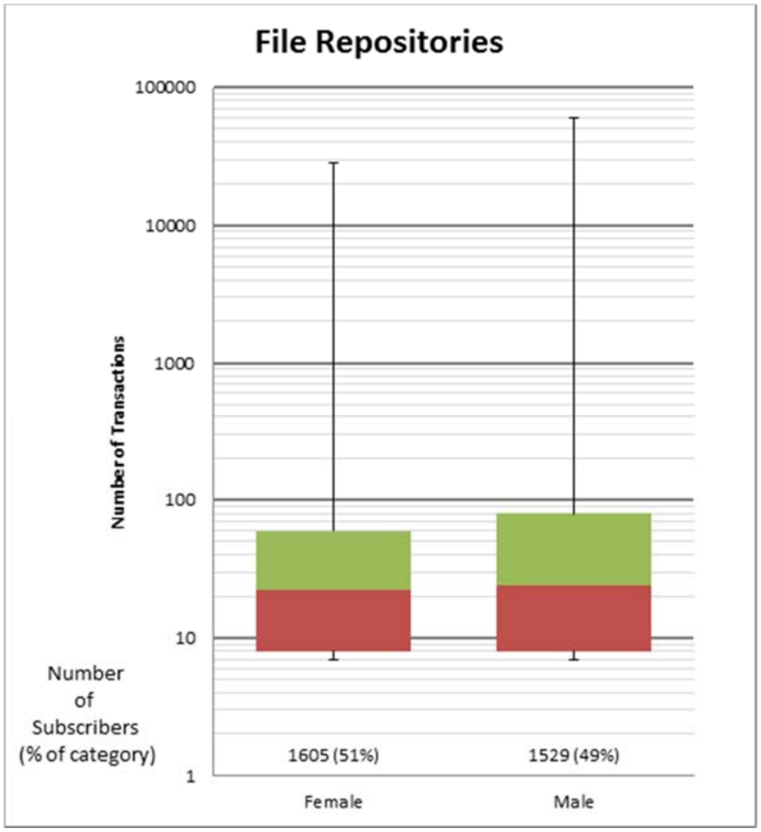

Content Server
1000000
100000
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
3582 (49%)
3747 (51%)
Female
Male

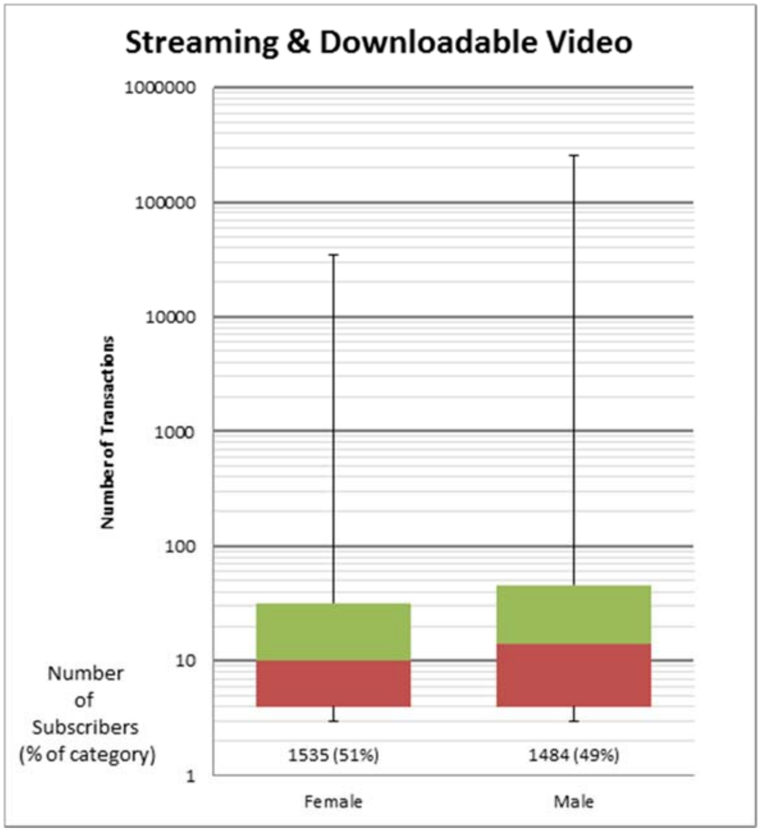

Games
1000000
100000
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
2224 (53%)
1992 (47%)
Female
Male

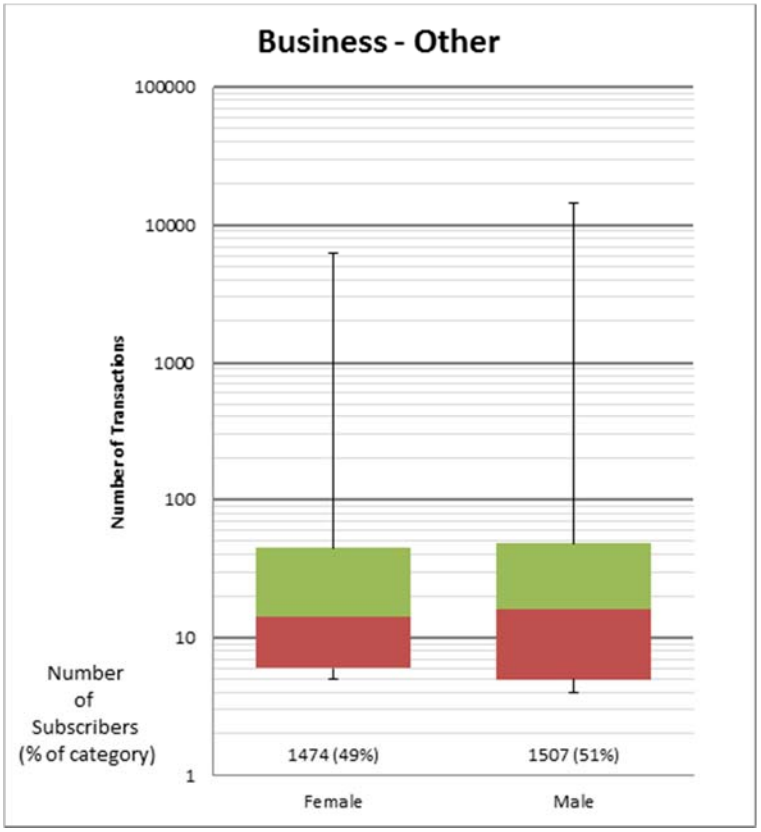

News
100000000
10000000
1000000
100000
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
1459 (44%)
1830 (56%)
Female
Male

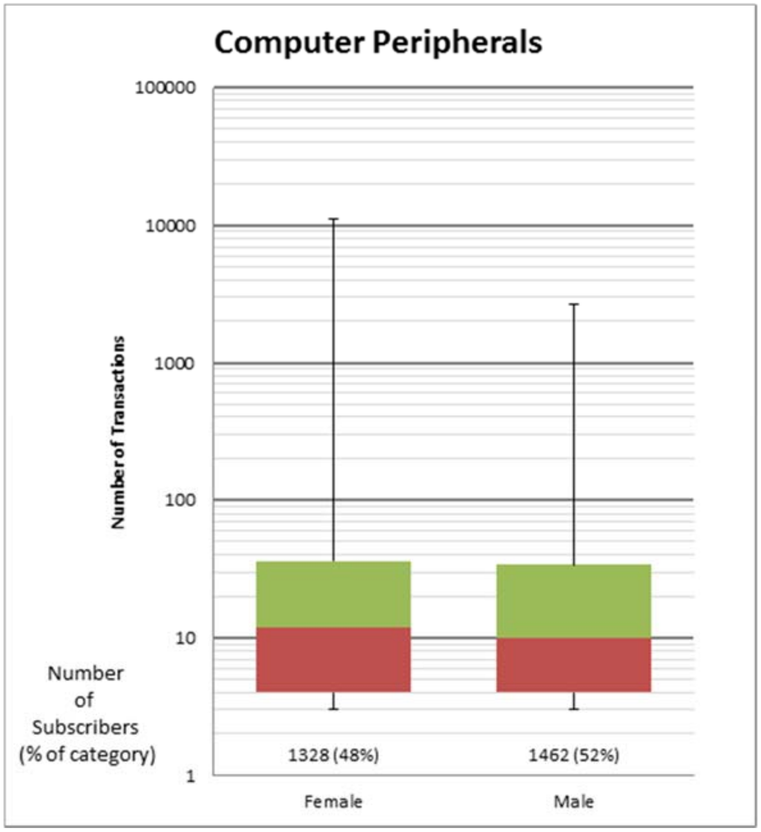

Portal Sites
100000
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
1746 (53%)
1533 (47%)
Female
Male

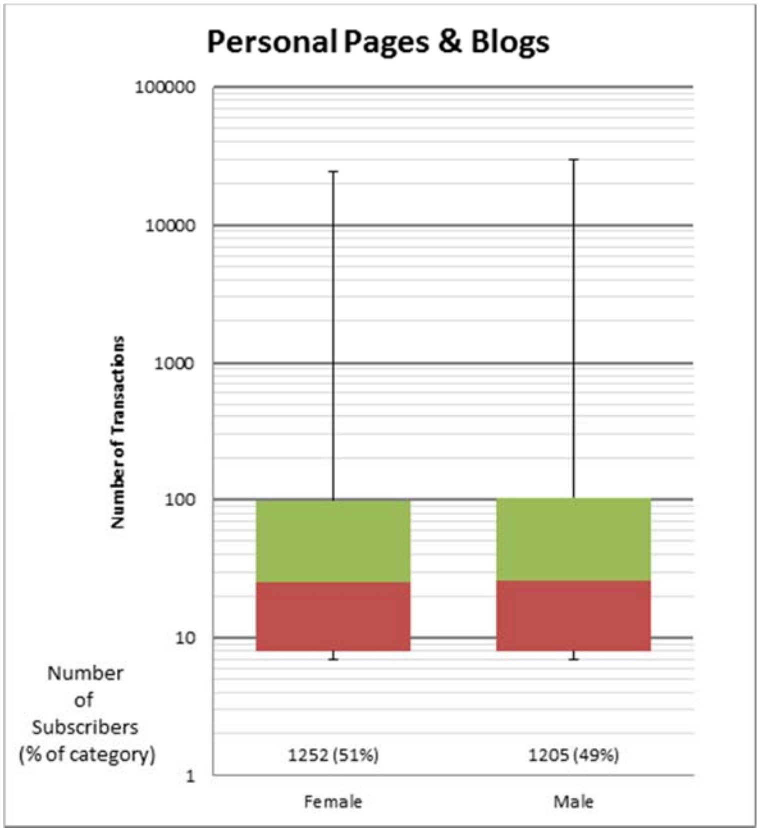

Information Security
Number of Transactions
100000
10000
1000
100
10
1
Number
of
Subscribers
(% of category)
1462 (46%)
1721 (54%)
Female
Male

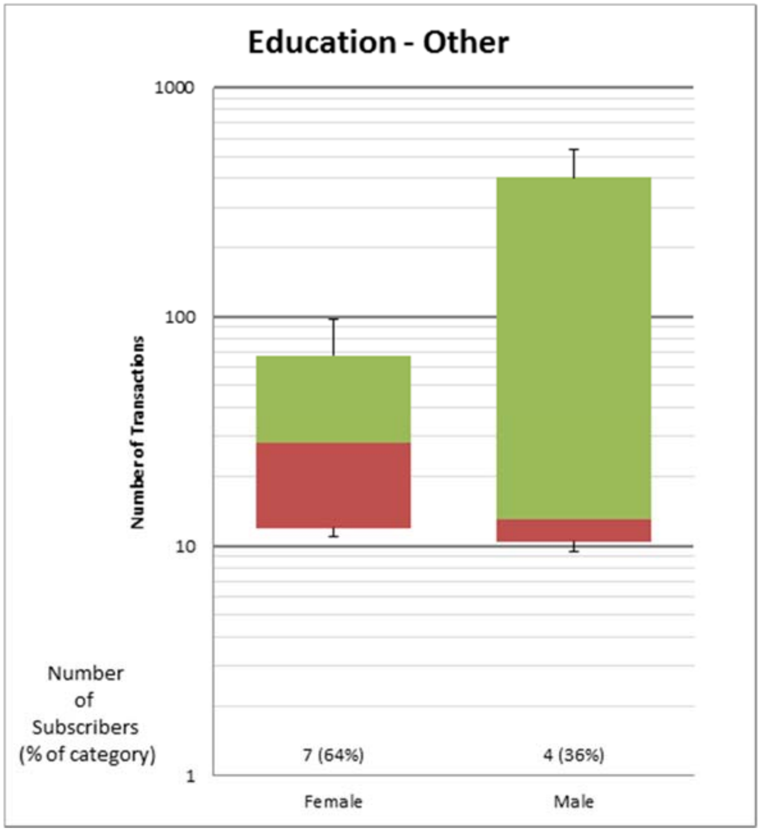

File Repositories
Number of Transactions
100000
10000
1000
100
10
1
Number
of
Subscribers
(% of category)
1605 (51%)
1529 (49%)
Female
Male

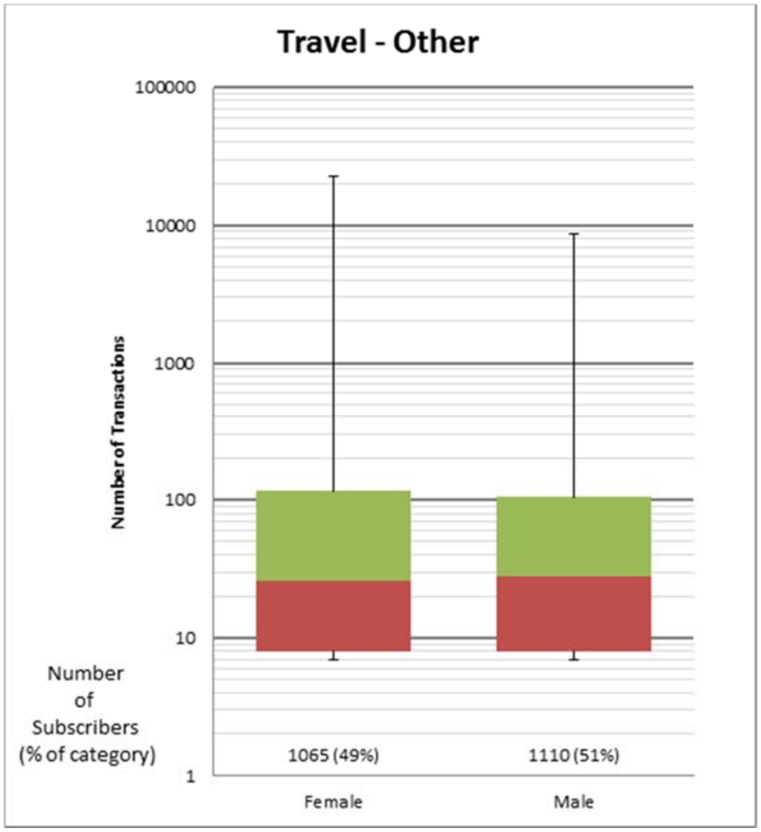

Streaming & Downloadable Video
1000000
100000
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
1535 (51%)
1484 (49%)
Female
Male

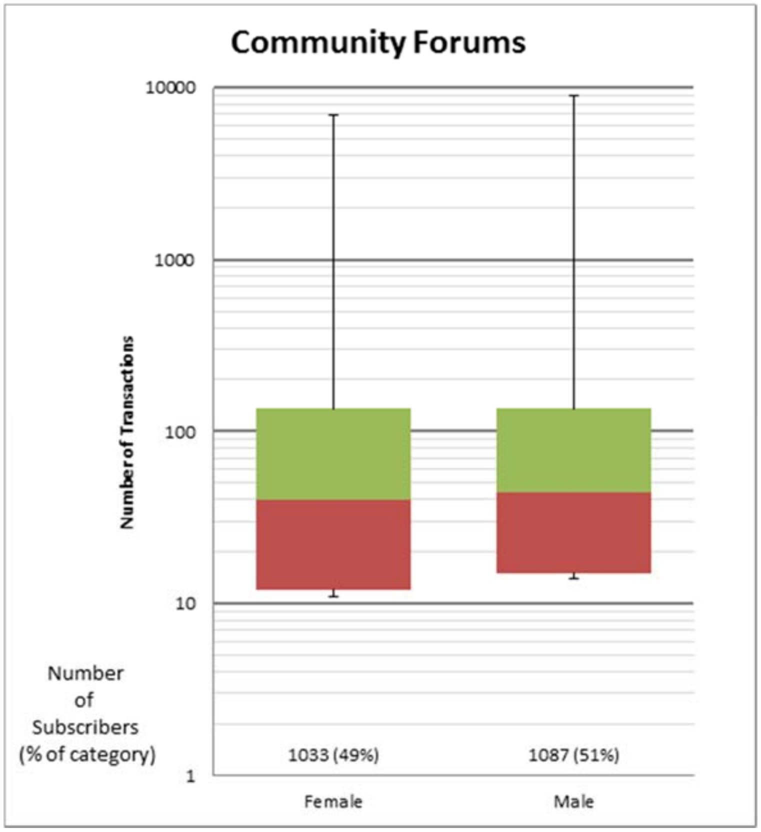

Business - Other
100000
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
1474 (49%)
1507 (51%)
Female
Male

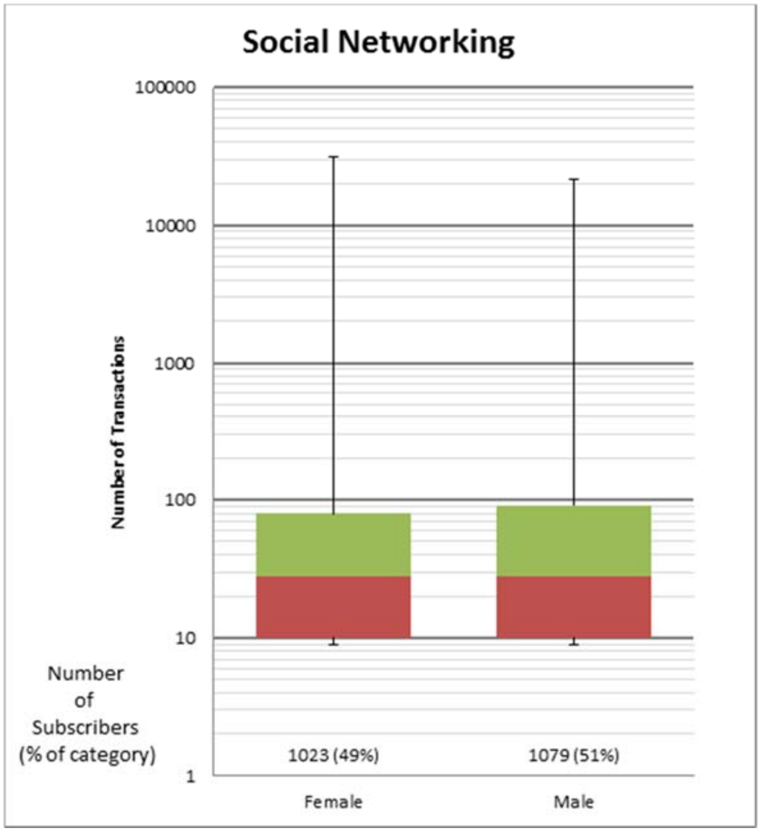

Computer Peripherals
100000
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
1328 (48%)
1462 (52%)
Female
Male

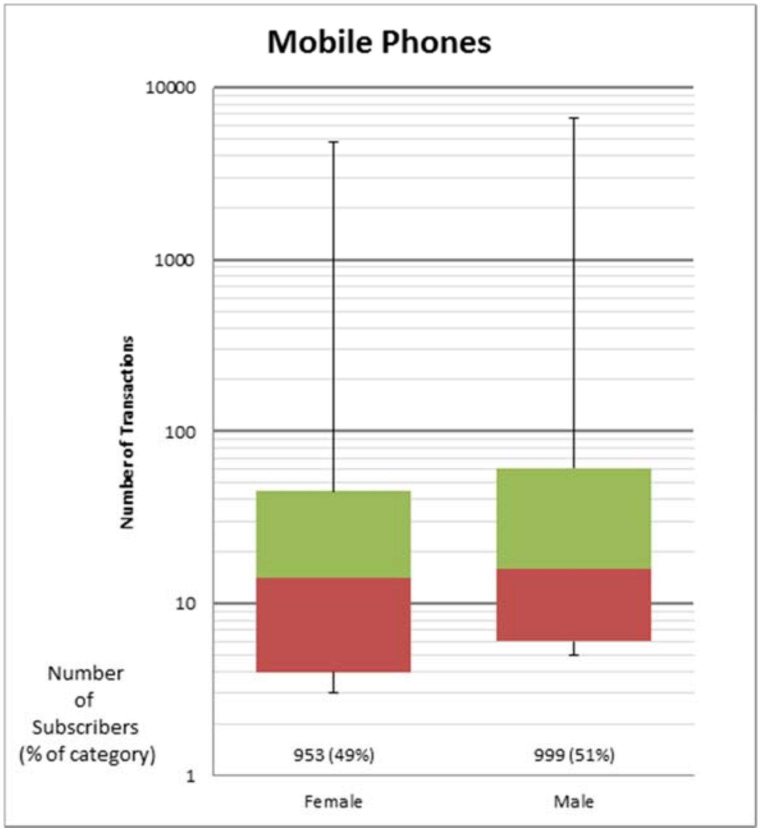

Personal Pages & Blogs
100000
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
1252 (51%)
1205 (49%)
Female
Male

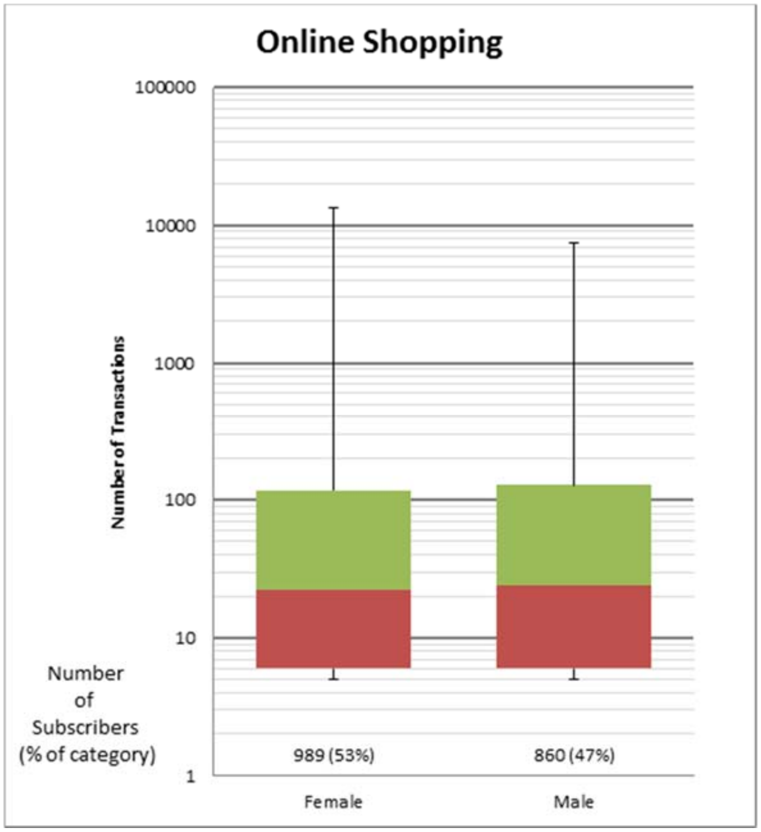

Education - Other
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
7 (64%)
4 (36%)
Female
Male


Travel - Other
100000
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
1065 (49%)
1110 (51%)
Female
Male


Community Forums
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
1033 (49%)
1087 (51%)
Female
Male


Social Networking
100000
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
1023 (49%)
1079 (51%)
Female
Male

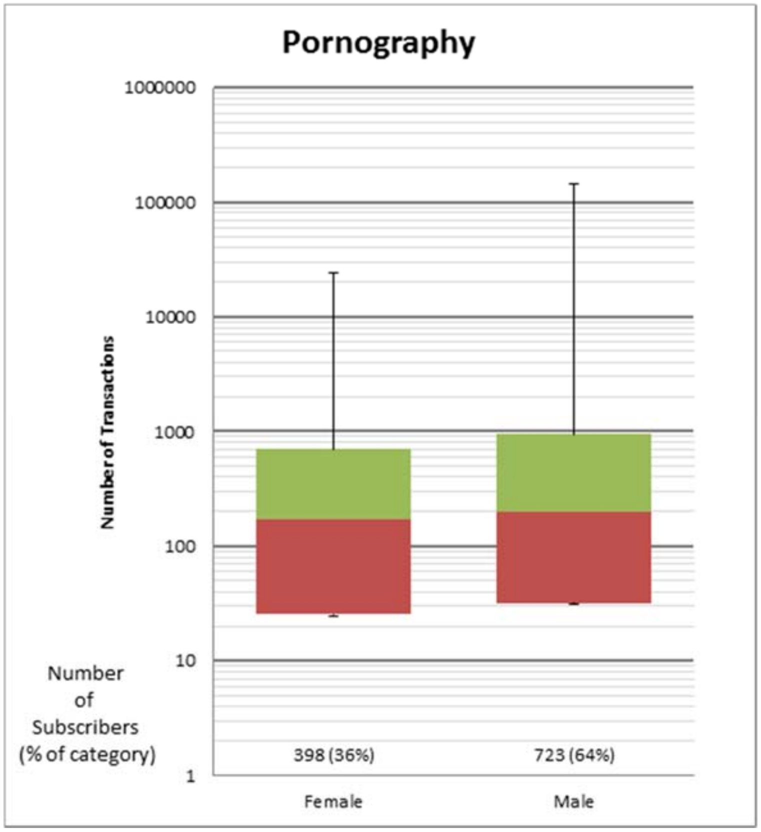

Mobile Phones
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
953 (49%)
999 (51%)
Female
Male

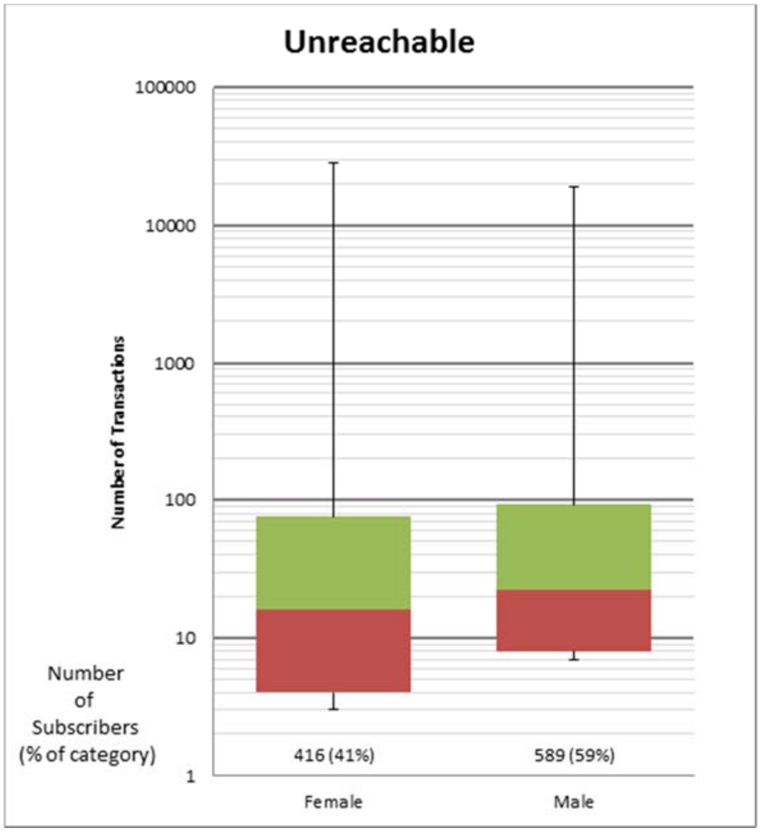

Online Shopping
100000
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
989 (53%)
860 (47%)
Female
Male

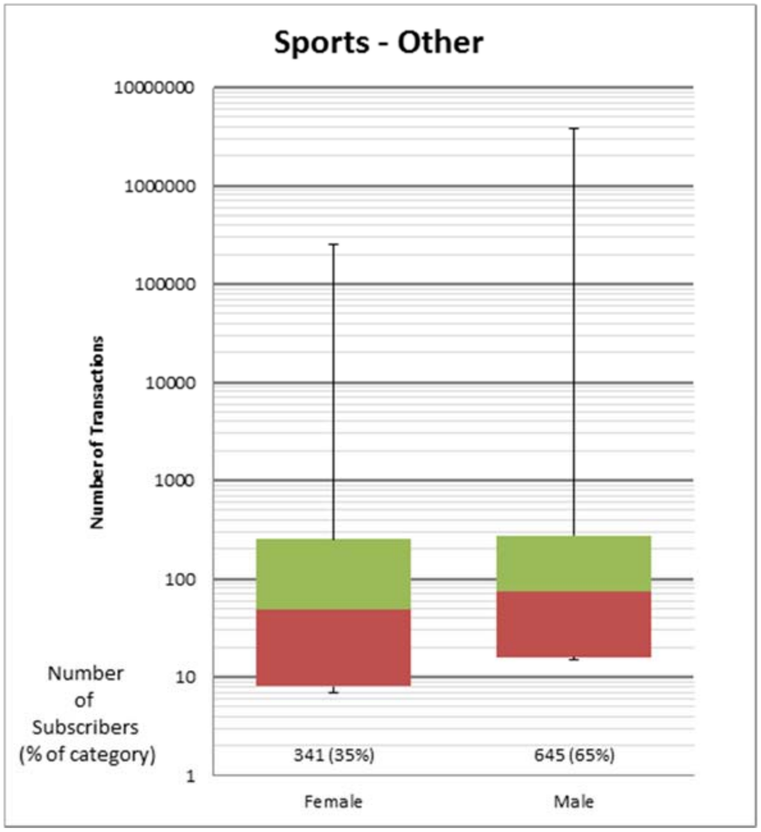
10 APPENDIX E: BOX POLOTS FOR TOP 20 DIFFERENTIATING CATEGORIES

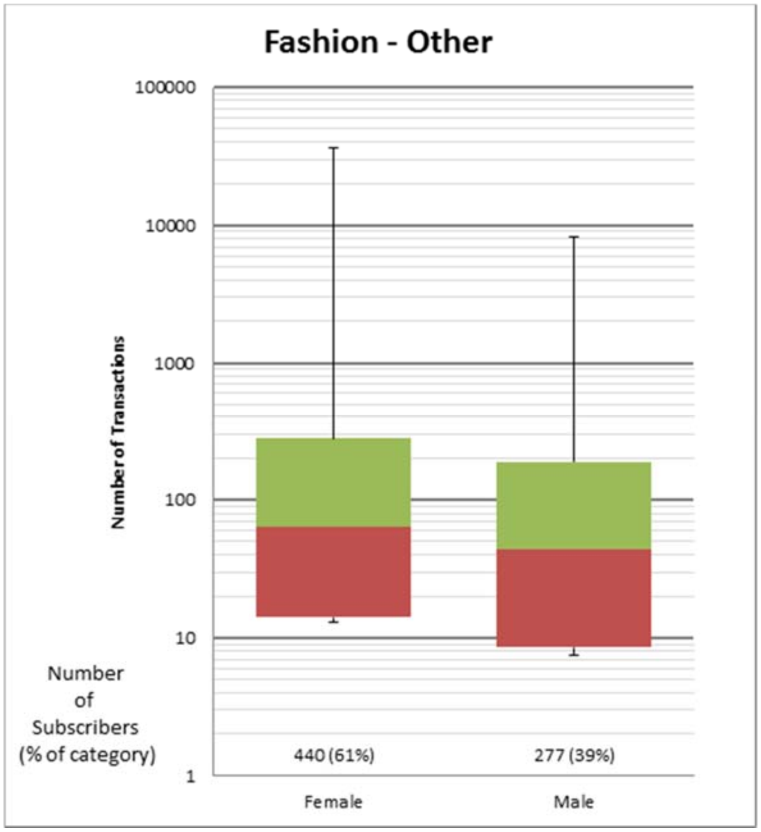
The red and green coloring of the box plots were a function of the graphics software we used to plot the data, and do not hold any significance other than differentiating the middle quartiles.

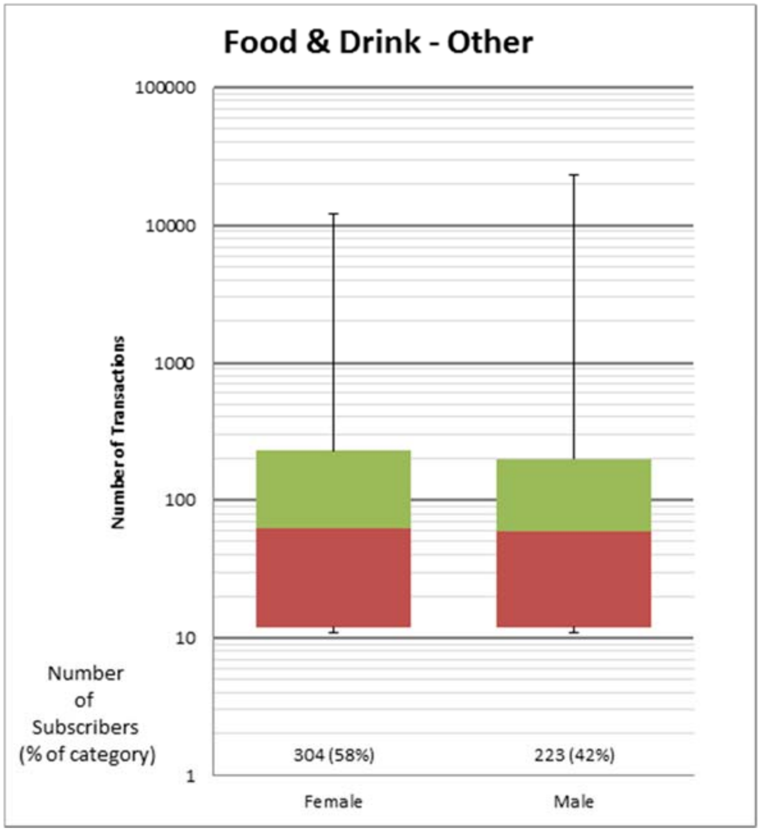
Rank	Category
25	Pornography
27	Unreachable
29	Sports - Other
33	Fashion - Other
40	Food & Drink - Other
41	Gambling
43	Coupons
45	Auctions & Marketplaces
47	Streaming & Downloadable Audio
48	Radio
53	Arts - Other
54	Dating & Relationships
58	Malware Distribution Point
59	Educational Institutions
59	R-Rated
61	Cartoons & Anime
62	Instant Messenger
63	Personal Storage
64	Piracy & Copyright Theft
65	Humor
67	Sex & Erotic
68	Construction
73	Legal Issues
74	Hobbies & Interests - Other
74	Product Reviews & Price Comparisons
76	Gay
77	Chat
78	Home & Garden - Other

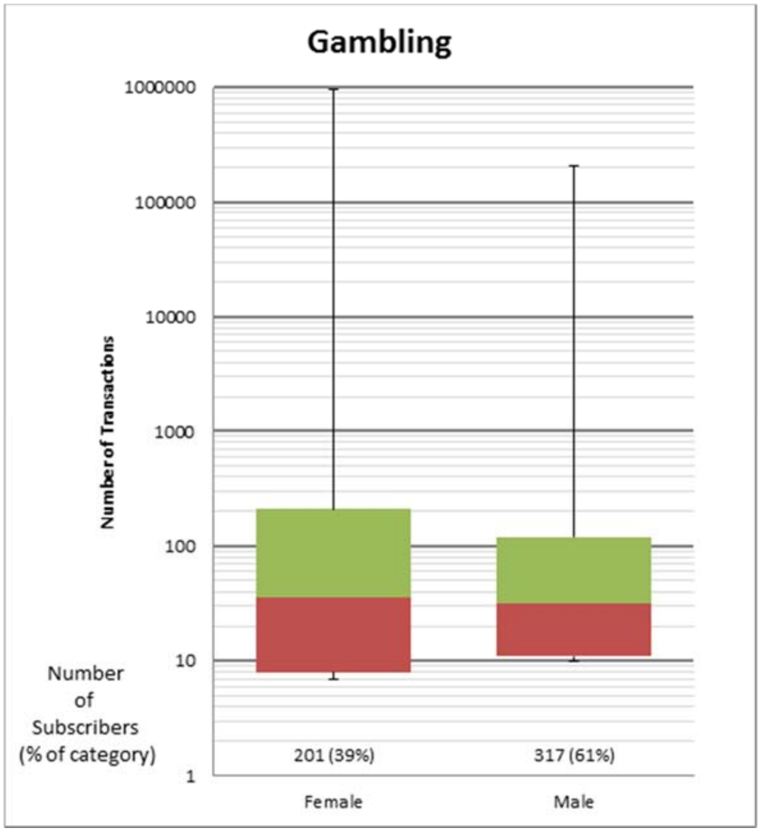

Pornography
1000000
100000
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
398 (36%)
723 (64%)
Female
Male

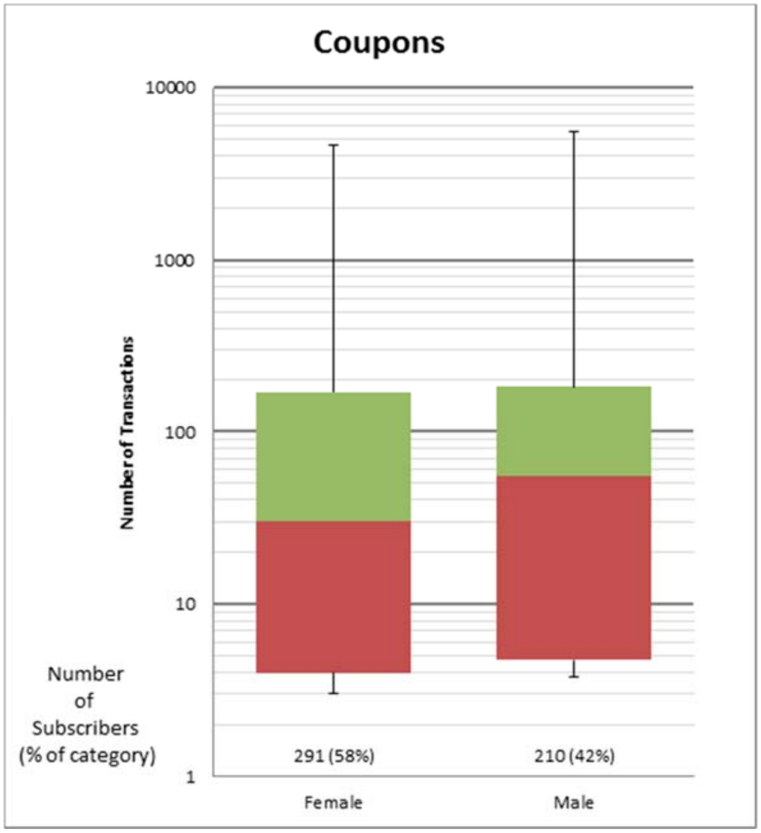

Unreachable
100000
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
416 (41%)
589 (59%)
Female
Male

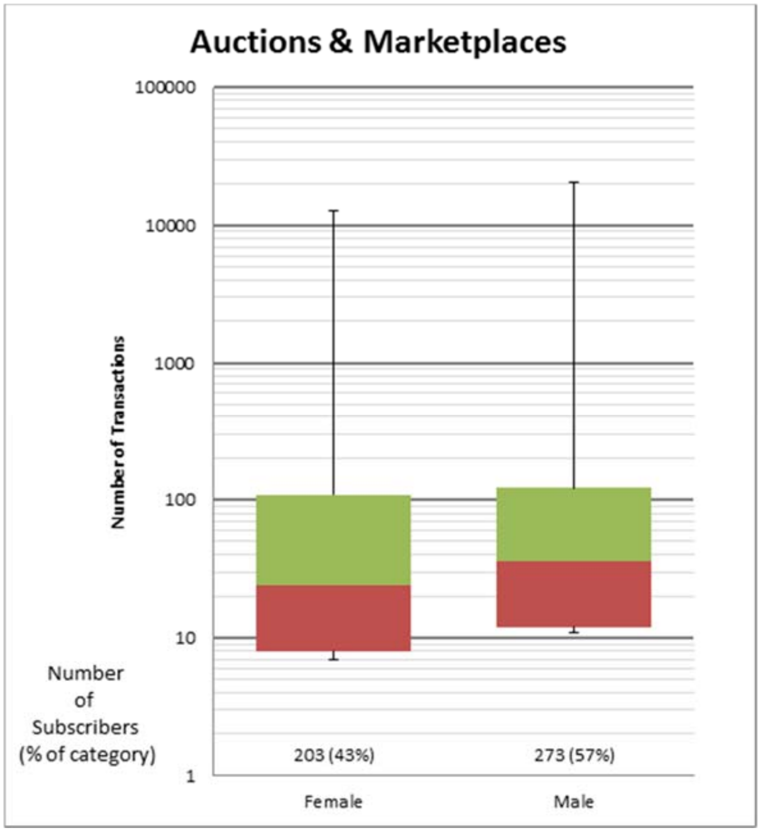

Sports - Other
10000000
1000000
100000
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
341 (35%)
645 (65%)
Female
Male

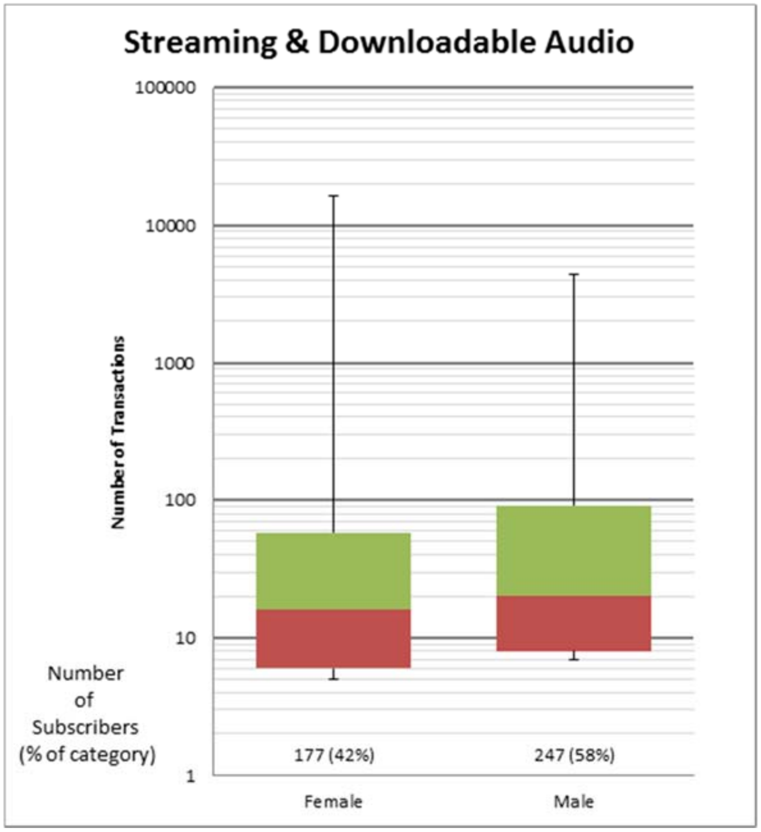

Fashion - Other
100000
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
440 (61%)
277 (39%)
Female
Male

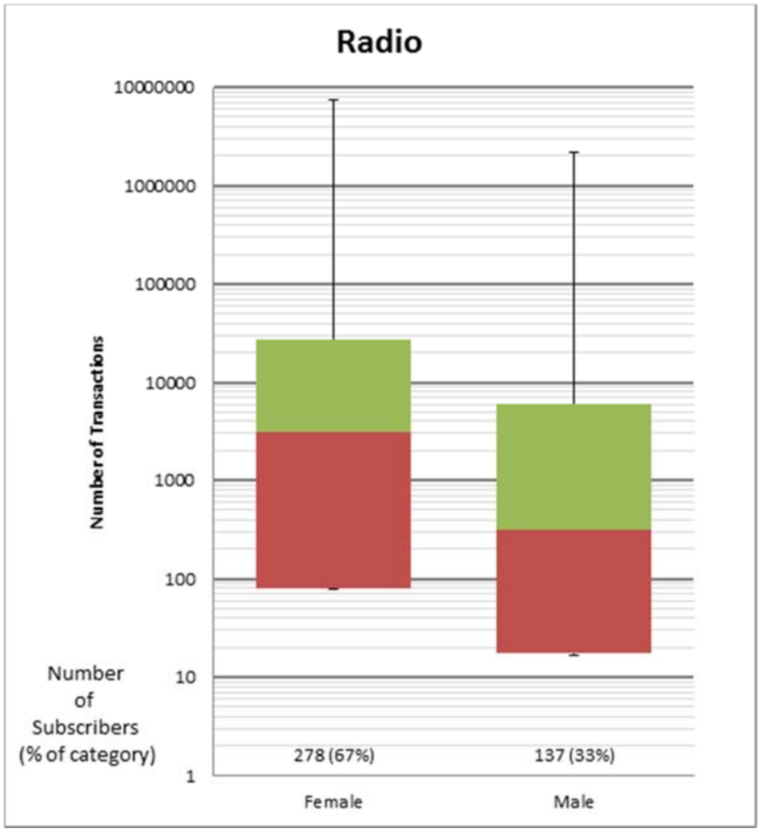

Food & Drink - Other
100000
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
304 (58%)
223 (42%)
Female
Male

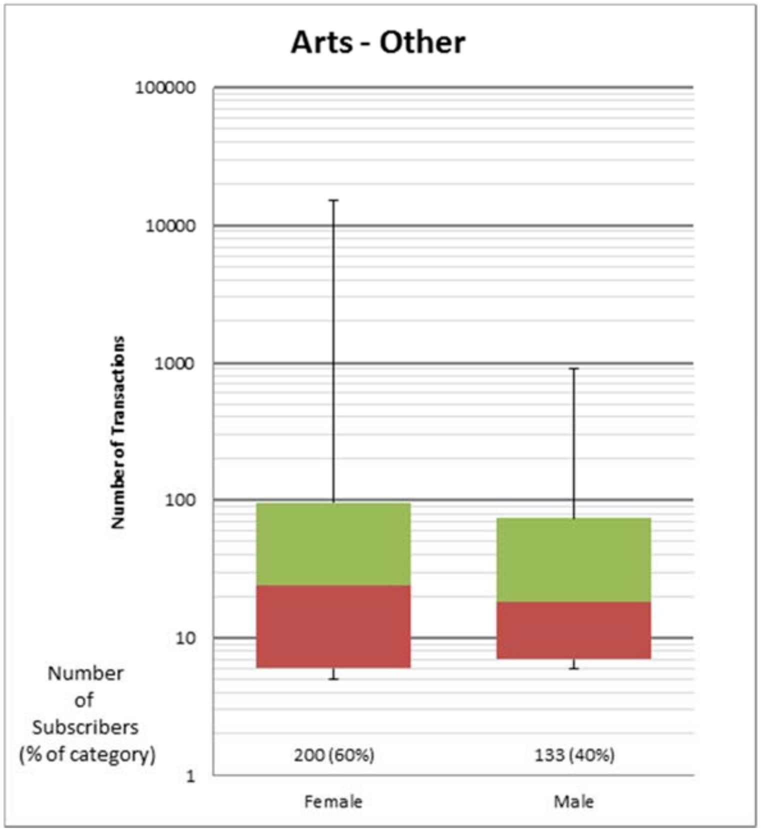

Gambling
1000000
100000
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
201 (39%)
317 (61%)
Female
Male

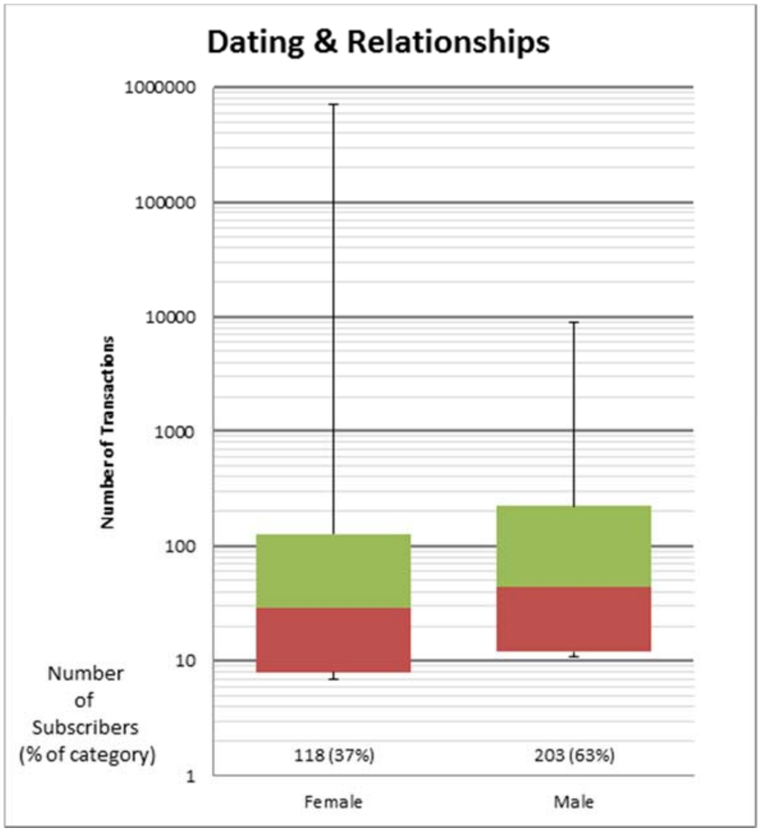

Coupons
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
291 (58%)
210 (42%)
Female
Male

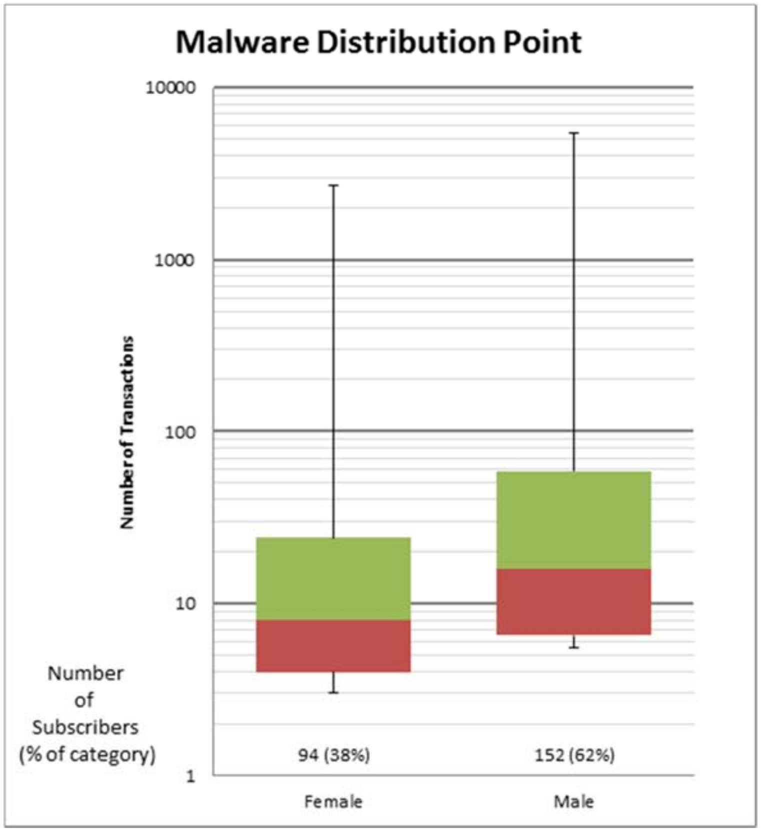

Auctions & Marketplaces
100000
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
203 (43%)
273 (57%)
Female
Male

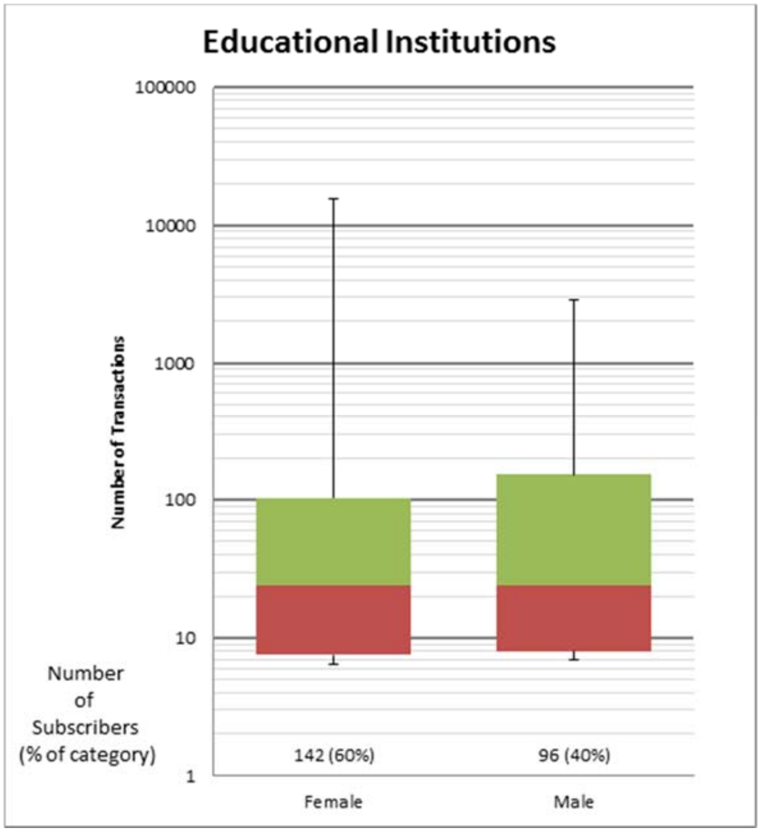

Streaming & Downloadable Audio
100000
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
177 (42%)
247 (58%)
Female
Male

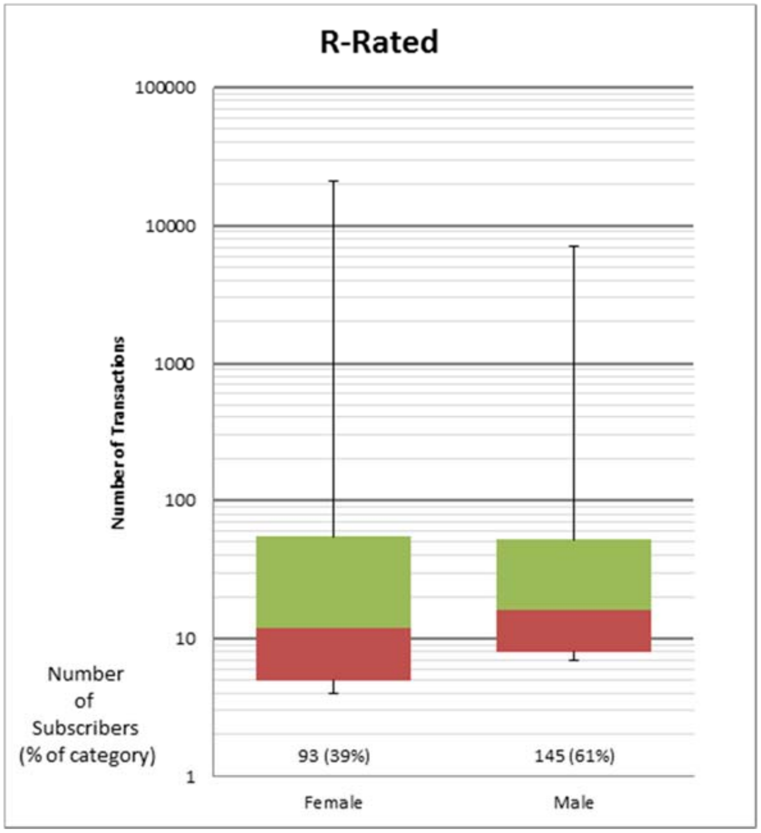

Radio
10000000
1000000
100000
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
278 (67%)
137 (33%)
Female
Male

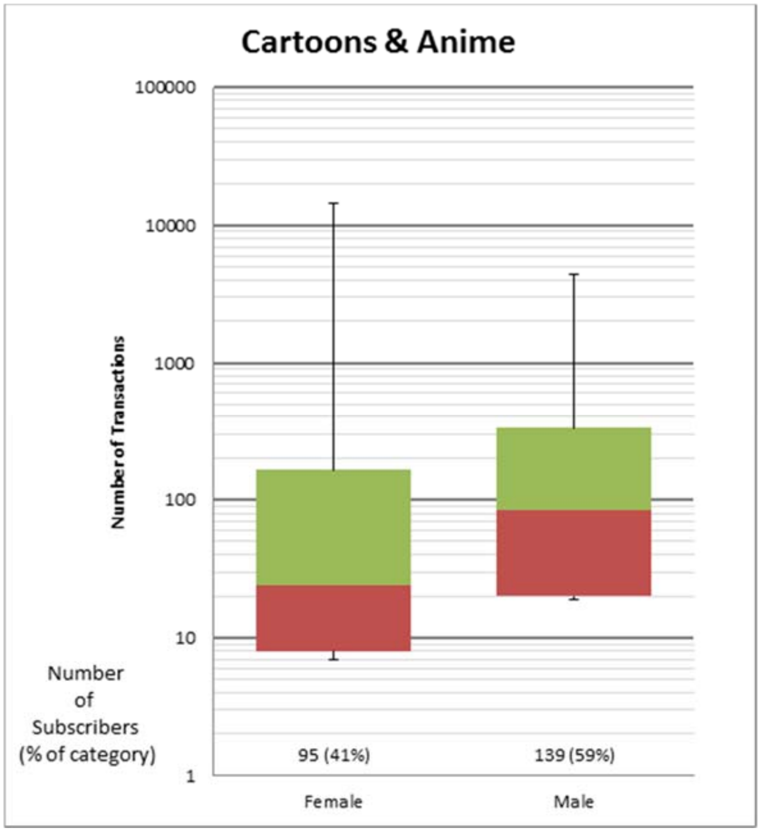

Arts - Other
Number of Transactions
100000
10000
1000
100
10
1
Number of Subscribers (% of category)
200 (60%)
133 (40%)
Female
Male

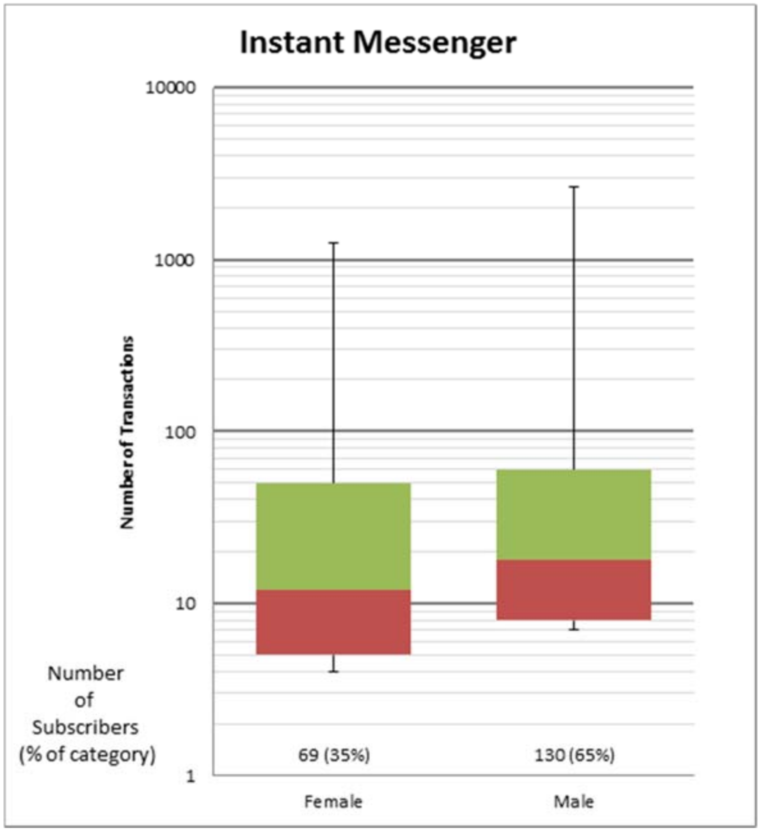

Dating & Relationships
Number of Transactions
1000000
100000
10000
1000
100
10
1
Number of Subscribers (% of category)
118 (37%)
203 (63%)
Female
Male

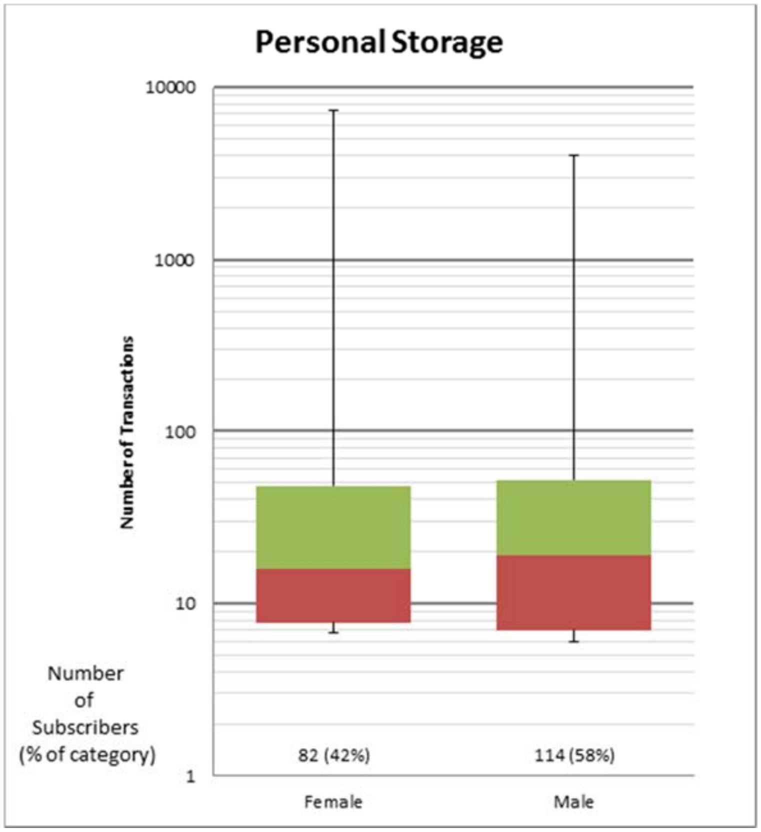

Malware Distribution Point
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
94 (38%)
152 (62%)
Female
Male

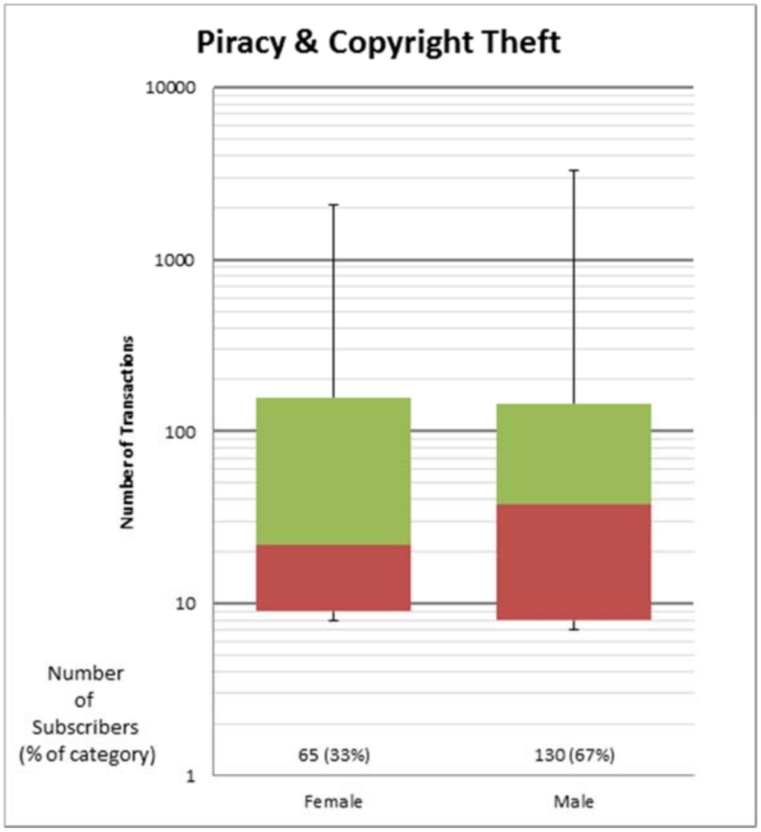

Educational Institutions
100000
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
142 (60%)
96 (40%)
Female
Male

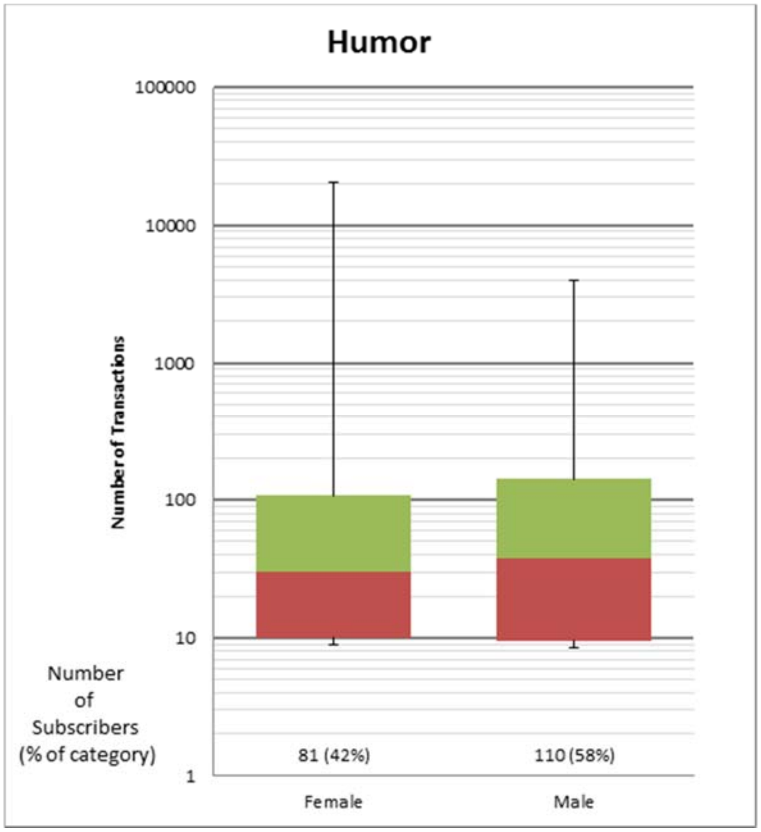

R-Rated
100000
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
93 (39%)
145 (61%)
Female
Male

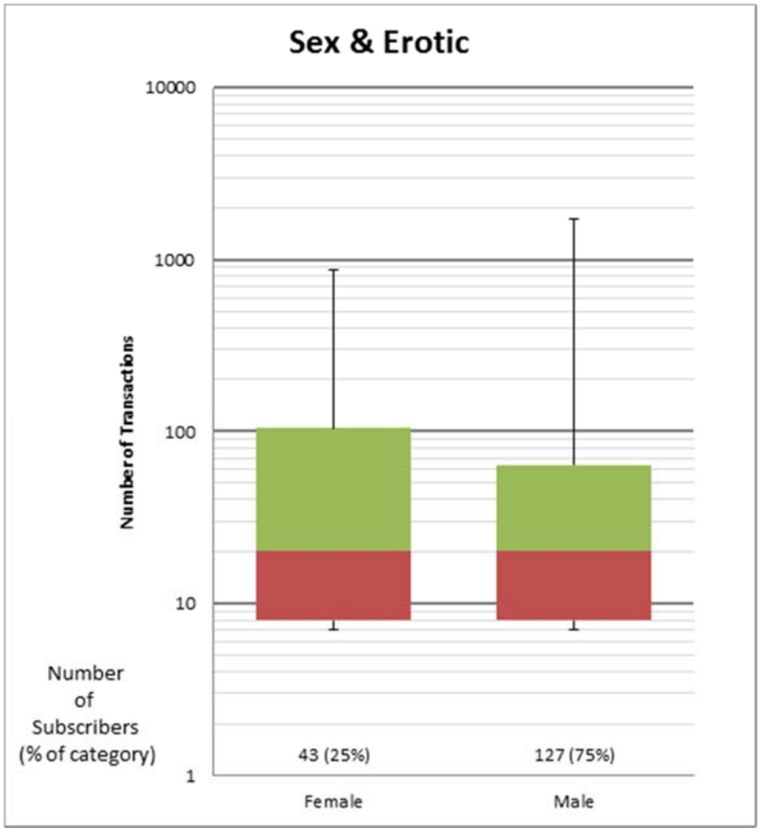

Cartoons & Anime
100000
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
95 (41%)
139 (59%)
Female
Male

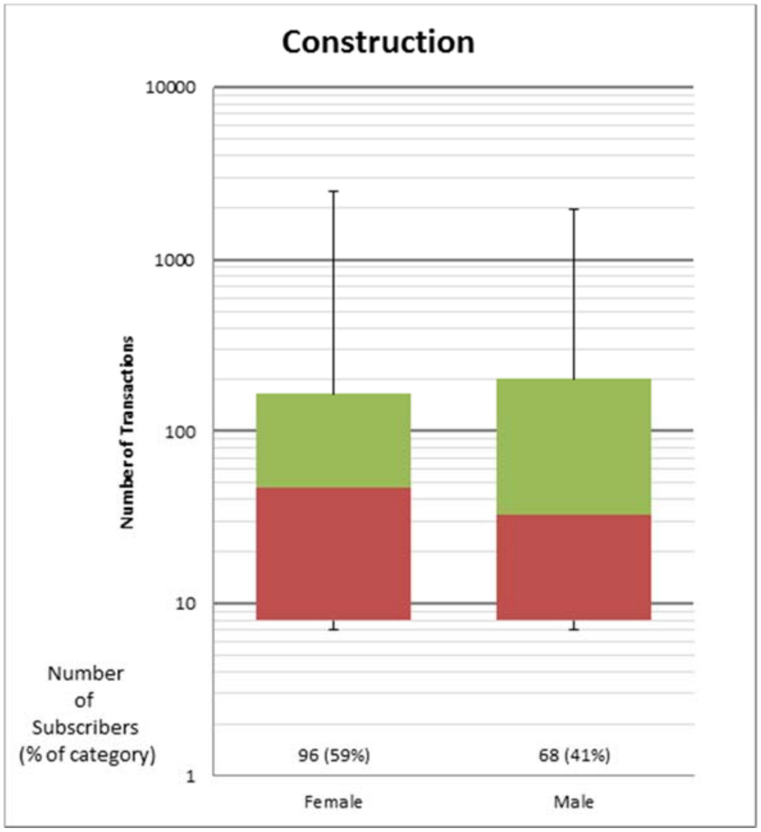

Instant Messenger
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
69 (35%)
130 (65%)
Female
Male

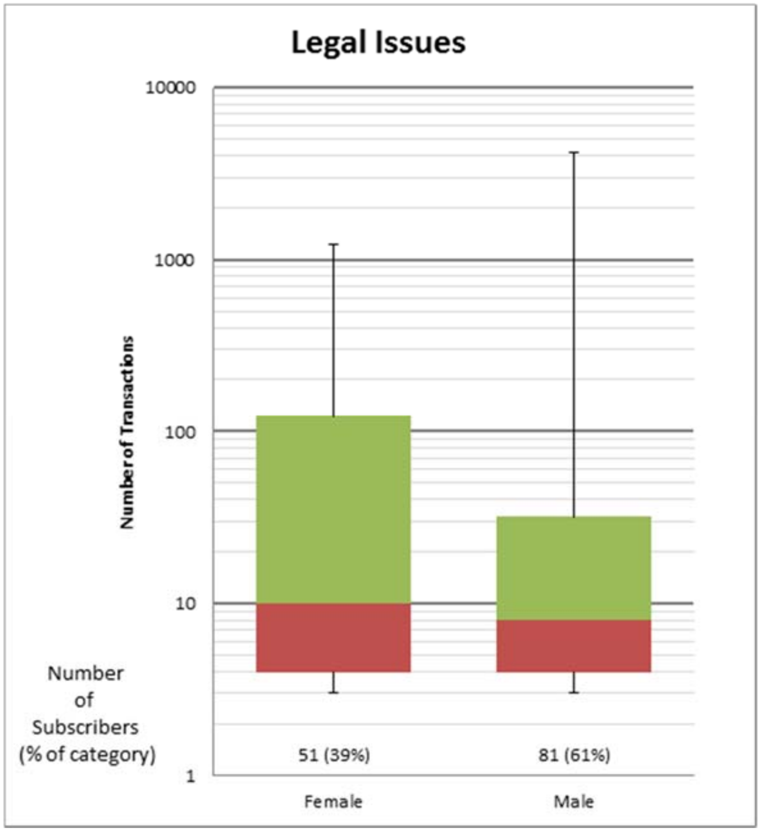

Personal Storage
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
82 (42%)
114 (58%)
Female
Male

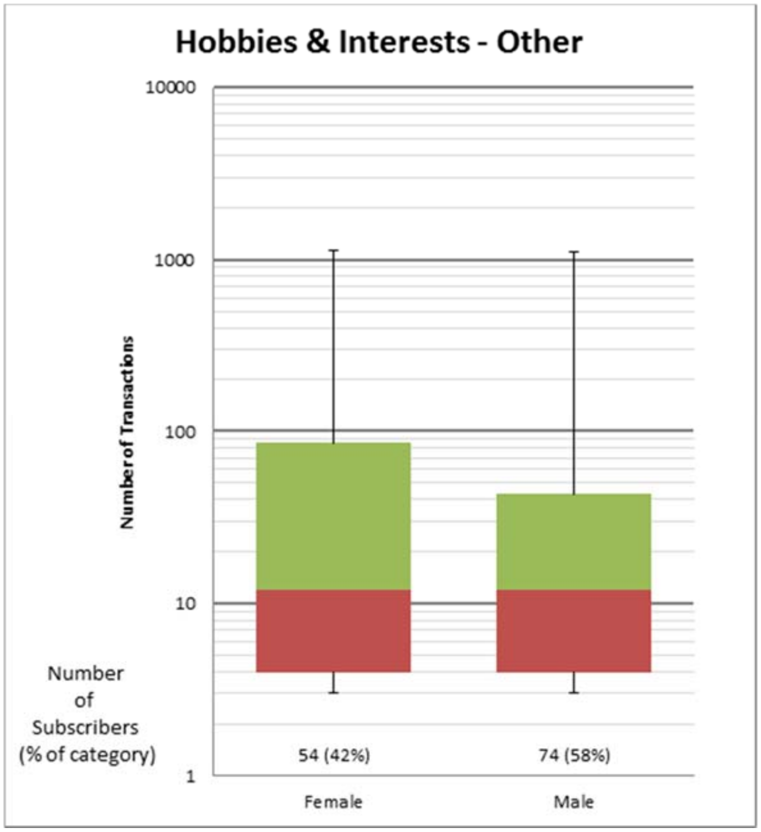

Piracy & Copyright Theft
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
65 (33%)
130 (67%)
Female
Male

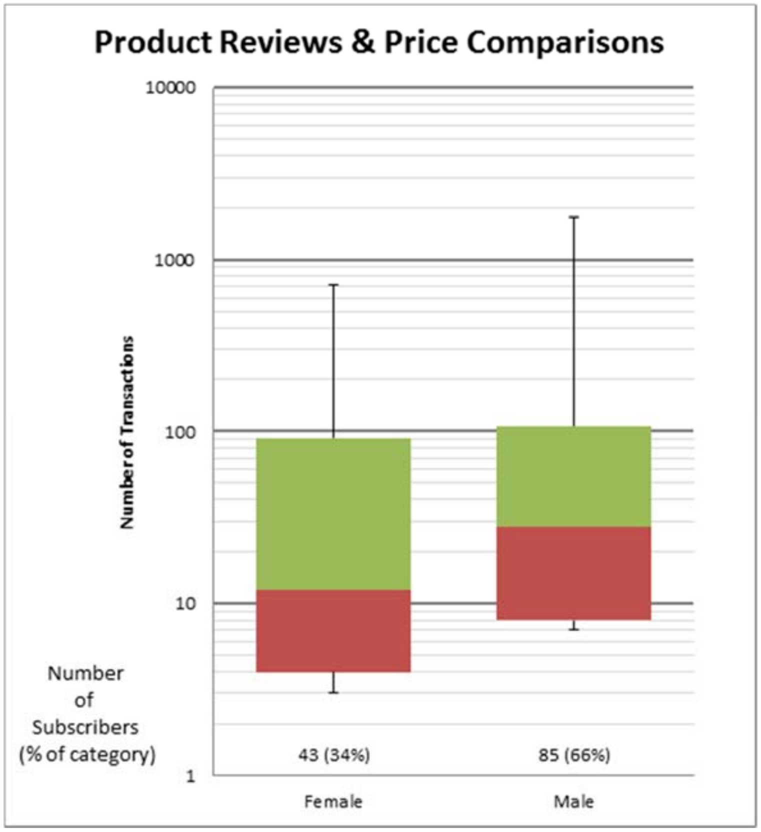

Humor
100000
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
81 (42%)
110 (58%)
Female
Male

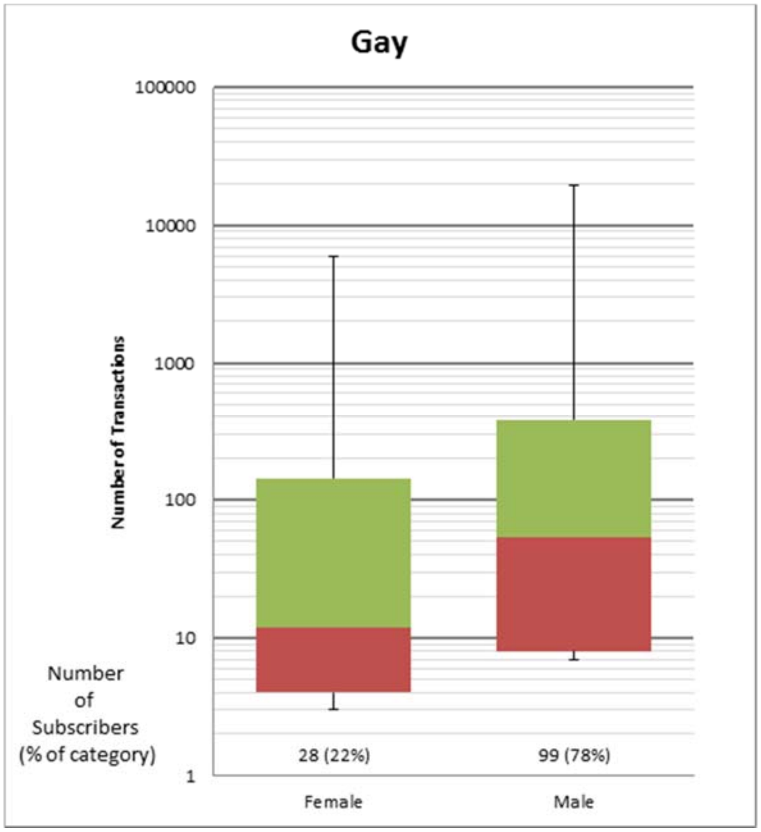

Sex & Erotic
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
43 (25%)
127 (75%)
Female
Male

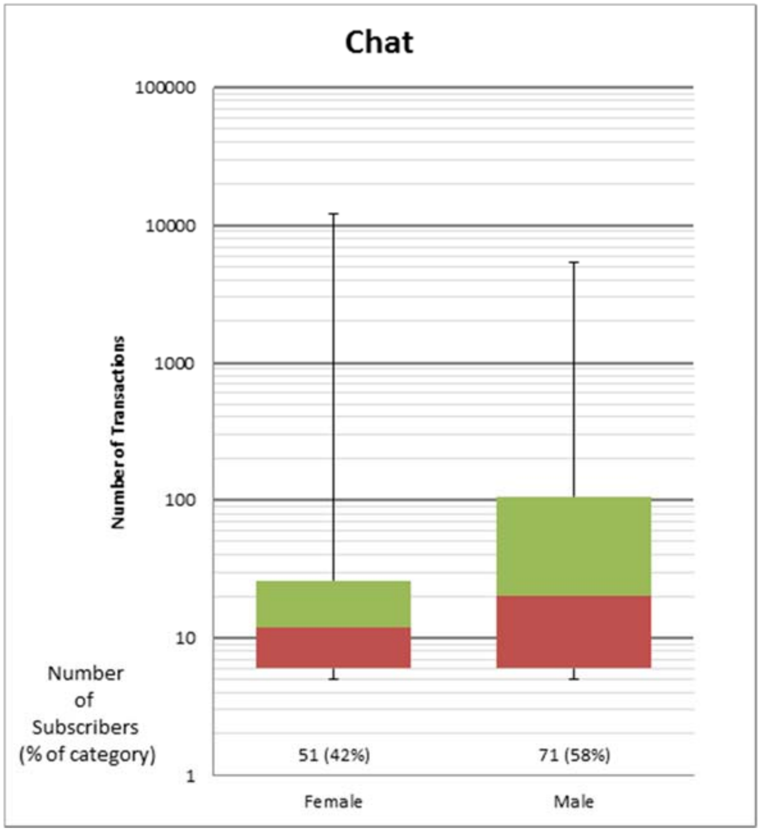

Construction
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
96 (59%)
68 (41%)
Female
Male

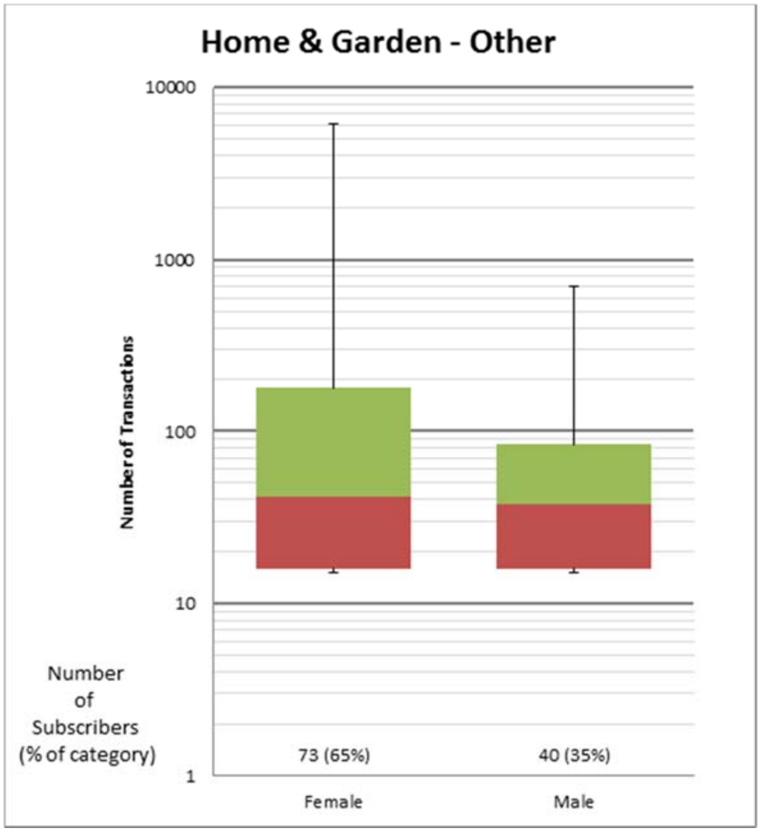

Legal Issues
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
51 (39%)
81 (61%)
Female
Male


Hobbies & Interests - Other
10000
1000
100
10
1
Number of Transactions
Number of Subscribers (% of category)
54 (42%)
74 (58%)
Female
Male


Product Reviews & Price Comparisons
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
43 (34%)
85 (66%)
Female
Male


Gay
100000
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
28 (22%)
99 (78%)
Female
Male


Chat
100000
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
51 (42%)
71 (58%)
Female
Male


Home & Garden - Other
10000
1000
100
10
1
Number of Transactions
Number
of
Subscribers
(% of category)
73 (65%)
40 (35%)
Female
Male