



Advancing Financial Inclusion

# Guidance Note 2: A technical guide to FinNeeds data analysis

June 2019

Established and powered by



In collaboration with





# Contents

1	Analysing FinNeeds demand-side survey data	2
2	Retrofitting the FinNeeds framework on existing datasets	6
3	Transactional data analysis	10
4	Merged demand and transactional data analysis	14

## About insight2impact

---

insight2impact is a resource centre that aims to catalyse the provision and use of data by private and public-sector actors to improve financial inclusion through evidence-based, data-driven policies and client-centric product design.

insight2impact is established and powered by Cenfri and FinMark Trust. The programme is funded by Bill & Melinda Gates Foundation in partnership with The MasterCard Foundation.

### For more information:

Visit our website at [www.i2ifacility.org](http://www.i2ifacility.org).

Email Mari-Lise du Preez (Partnerships Manager) at [mari-lise@i2ifacility.org](mailto:mari-lise@i2ifacility.org).

Call us on +27 21 913 9510.

The FinNeeds toolkit sets out two main data sources that can be used as a basis for compiling FinNeeds indicators: demand-side survey data and transactional data. How is it possible to analyse different data sources to extract actionable insights suitable to a specific context?

This technical guide expands the guidance provided in the FinNeeds toolkit on the data analysis steps and methods. It applies the FinNeeds analytical framework to any of the following data sources as outlined in the Toolkit:

1. FinNeeds survey data
2. Other demand-side survey datasets not designed according to the FinNeeds framework
3. Transaction databases
4. Merged datasets containing the same set of consumers generated via a demand-side survey as well as extracted from transactional data

# 1 Analysing FinNeeds demand-side survey data

## Step 1: Getting familiar with the structure of the data

A dedicated FinNeeds questionnaire delivers a flat table, or a cross-sectional dataset with no time dimension, as does the incorporation of a FinNeeds module or questions into a host survey. These surveys can have hundreds up to a couple of thousand columns and rows:

- Variables, derived from the questions, are captured as columns
- Observations (individuals) are captured as rows

Most datasets contain one or more weights that allow the analysis to be expanded to be representative of the overall population (or for a specific state or locality, depending on the sampling methodology) at individual or household level.

## Step 2: Cleaning the data

**Recoding selected variables.** Demand-side survey questionnaires are usually implemented by a local research house. This means that the datasets that are generated may have some differences in the way that the data is captured. For example, multiple-choice questions and their responses can be captured in several different ways. Depending on what the researcher would like to achieve, the resulting variables may need to be recoded

to create variables that are easier to manipulate. For example, in some datasets multiple choice questions are captured using binary variables that take the value zero when the respondent answers “No” but take different values (sequentially or not) for “Yes”. Should you wish to use these binary variables to construct an aggregate variable, such as a binary variable for whether the respondent uses a formal product, then it may be necessary to recode all of the binary variables so that they equal one when the respondent answers “Yes” – when working in many statistics programmes, this is the easiest way to proceed. There are many of these and other types of data manipulation required before a particular dataset can be used to investigate questions of interest.


**Sense-checking the data.** Apart from recoding variables as needed, the researcher should conduct an overall sense check on the data to ensure it matches the questionnaire and the answers align with the options provided in the questionnaire. Data can be sense-checked and cleaned across a number of dimensions:

- **Checking for data gaps and ambiguities.** Understanding the exact meaning of all responses is necessary to determine the appropriate denominator for percentages. Pertinent questions to ask as part of the data cleaning process, include: Are there responses that are coded ambiguously?



Most datasets contain one or more weights that allow the analysis to be expanded to be representative of the overall population [...] at individual or household level.





Are some responses assigned a value that has no description or does not exist in the questionnaire and was not explained by the research firm? If so, these variables should be removed.

- **Sense-checking demographic data:**

Demographic variables should all be checked and cleaned before use. For example:

- A question on gender should result in an approximate 50:50 split between the genders, unless there is a contextual reason to expect otherwise.
- Are there people who are older than, say, 120 or younger than 16 when the questionnaire should only have been asked of adults aged 16 and older? If so, such outliers should be removed.

- **Sense-checking against economic intuition and existing data sources.** All variables require a sense check by asking if the results confirm economic intuition. This can be done by tabulating the main variables. Results that differ significantly from what economic theory or intuition would suggest, should be treated with caution. Potential checks include:

- Does the income data, if available, follow a normal income distribution?
- Do statistics, such as the number of rural adults, education distribution or age distribution, match equivalent figures from other reliable and relevant datasets available?

- Standard statistical techniques for the identification of outliers may be useful, but outliers should be rare.

## Step 3: Conducting the analysis

**Analytical framework.** It is important, at the outset, to plan the analysis according to a clear analytical framework informed by the FinNeeds conceptual framework as set out in the toolkit. The core elements of the FinNeeds framework as set out in the Toolkit forms the basis for the analytical framework, namely:

1. **Use case incidence** for each need category
2. **Devices used towards each need category**, as well as for key use cases in each category, labelled according to the device taxonomy as outlined in the toolkit's main text and Guidance Note 1: Demand-side data collection guidelines:

- **Product dimension:** according to this dimension, devices can be labelled as either credit, savings, payments, insurance or assistance (a term we use to refer to non-reciprocal financial support from the respondent's social circle or community)
- **Provider dimension:** according to this dimension, devices can be labelled as either formal, informal, social or personal devices<sup>1</sup>

3. **Usage analysis** for specific types of financial devices as relevant – exploring the frequency, recency, monetary value and duration dimensions of usage as relevant to the type of financial device.

---

<sup>1</sup> Formal devices (services or products) are those offered by regulated financial institutions such as banks. Informal devices are offered by unregulated businesses and individuals. Social devices are those that rely on social networks such as family and friends or cooperative savings associations. Personal devices are mechanisms that individuals employ themselves, such as saving in cash in a hiding place or reducing expenditure.



**4. Drivers of use:** by considering the questions on self-reported reasons of use and categorising those into functional, relational and personal characteristics, as explained in the toolkit.

**5. Outcomes of use:** by, for example, attempting to create “treated” groups who are liquid, or were able to recover from a shock and to establish in what ways they differ from “untreated” groups. If possible, the ideal is to devise a test to establish which factors cause individuals to fall within the treatment group: why are they liquid or resilient?

**General investigation.** The first step when starting to analyse the data is to look at the unweighted and weighted tabulations of each main need category: use cases as well as devices used towards each need category. This can be done by exporting the tabulations to an external document, such as an Excel sheet, which allows for easy inspection. It is a worthwhile exercise to gain an overall sense of the financial needs landscape in the target country.

**Constructing need and device category variables.** The heart of the analysis rests on investigating what FinNeeds use cases respondents express, as well as how they meet those use cases. That is, what types of devices do respondents use

towards each need category or key use cases within each category. To do so, the researcher must first calculate incidence of use cases within each need category. This involves creating a binary variable which captures a list of questions that determine whether an individual has a certain need. For example, all of the questions relating to liquidity should feed into the binary variable which indicates whether a respondent has experienced liquidity problems in the past 12 months or not. Next, the researcher should construct variables to group devices used towards each need or use cases into the two key dimensions of the financial device taxonomy, namely:

- **Product dimension:** Is it a savings, credit, payments, insurance (where relevant) or assistance device? As with the binary variable for aggregate needs, the researcher can, for example, construct a binary variable that captures whether a respondent uses a savings product towards a resilience need, or a credit product towards the meeting goals need. Likewise, below.
- **Provider dimension:** Is it a formal, informal, social or personal device? The same method from product dimension can be applied with the provider dimension.

“

...the researcher must first calculate incidence of use cases within each need category [... which] involves creating a binary variable which captures a list of questions that determine whether an individual has a certain need.

”



**Segmenting target groups.** In exploring devices used towards needs, it will be important to explore key differences or skews between different groups of respondents. This is done by creating target groups or segments of interest based on the demographic variables included in the dataset. For example, many policymakers or market players may be interested in knowing how use case incidence and devices used towards needs manifest for specific population segments, such as women, youth, rural farmers or different socio-economic classifications. These segments can be explored through basic cross-tabulations.

More complex segments can be created using either a “designed clustering” or “unsupervised statistical clustering” method:

- **Designed clustering:** Respondents are categorised based on a key variable, such as their primary source of income. Income categories can be bundled based on similar characteristics, for example, whether it’s a formal source of income or according to regularity of income. Once income clusters have been identified, the profile of each group of respondents can then be compared in terms of demographics, use case incidence, device portfolio taken up towards various needs and use cases, as well as usage profiles.
- **Unsupervised clustering:** Various statistical clustering techniques or algorithms can be used to create clusters of respondents based on selected characteristics. For example, an indication of usage frequency, combined with certain demographic variables. This would entail allowing the algorithm to run and allocate individual respondents into different randomly selected clusters. In other words, the algorithm

runs various iterations until it has identified groups of individuals that are “different enough” statistically to be separated out. The device, usage or demographic profiles of such clusters can then be compared to derive FinNeeds insights. See [insert link to online guide on clustering] for a more technical explanation of how to conduct an unsupervised clustering exercise.

Once target group segments have been created, the segmentation can be applied to any of the elements of the analytical framework by cross-tabulating the various segments to the analytical framework element. For example, it can be used to show how use case incidences differ between rural and urban respondents or how the device mix towards a particular need differs between the genders or between higher or lower socio-economic groups.

**Generating insights.** Once the data has been properly set up and the relevant variables constructed, it should be a relatively simple task to complete the analysis and extract insights relevant to the different elements of the analytical framework. The analytical framework should be used as a basic structure to compile tables and graphs of relevant results, considering what socio-economic and demographic segmentations or cross-tabulations will be most relevant to demonstrate the findings.

See the toolkit for examples of how outputs from the analysis have been used in insight2impact’s pilot studies to populate each of the FinNeeds indicators. And how these, in turn, can be used to answer key policy and market strategy questions.

## 2 Retrofitting the FinNeeds framework on existing datasets

As explained in the toolkit, the FinNeeds lens can still be applied to analyse existing demand-side survey data, even if the dataset was not designed according to the FinNeeds framework or to incorporate any dedicated FinNeeds indicators. Such analyses will not have granularity or completeness of FinNeeds insights rendered by tailored FinNeeds datasets, but may nevertheless render useful policy insights. All results should be presented with the necessary caveats and disclaimers on methodological constraints due to the nature of the underlying data.

This section describes the methodology for analysing an existing financial inclusion demand-side survey database that was not specifically designed to gauge market dynamics from a needs perspective. It assumes the basic survey structure of the FinScope survey, but could also be constructed from demand-side survey databases with similar objectives and questions. The purpose is to obtain the insights available from the needs measurement framework without having to undertake a completely new survey.

### Step 1: Investigate the quality of the dataset and establish the ability to construct specific use cases

Prior to starting the exercise, the following should be assessed:

- Is the data granular enough to allow for this type of analysis?

- Is there sufficient qualitative and third-party (desktop) research to highlight which use cases are most prominent in the country?
- Is this research sufficient to justify the housing of most, if not all, of the survey respondents within the use cases?

If the answer is “yes” to all these questions, it will be possible to establish granular use cases. If not, then it may only be possible to get a sense of needs at an overarching level, or it may require qualitative and/or desktop research to better understand use cases.

### Step 2: Establish a use case taxonomy

Regardless of the outcome of the process outlined above, a taxonomy should be established of relevant questions from the survey questionnaire for each use case, and for how the use cases cluster into a need<sup>2</sup>. The taxonomy is important as it lays out all the available data that is relevant to the construction of a needs strand, in line with the principles outlined in the indicator section of this note. To construct the taxonomy, qualitative and quantitative research must be engaged side-by-side. The quantitative data will provide the information required to calculate the strand, while the qualitative information will help determine which use cases to include and consequently which use cases cluster together to form a need.

---

<sup>2</sup> A taxonomy is a repository of all the relevant questions that are available in a survey and that links them to specific themes that the researcher wishes to understand. In the case of the current measurement framework, these themes will be the relevant use cases established in the first step of the process.



## What does a use case taxonomy look like?

A taxonomy is built around the variables that are available to measure use cases categorising under each need. To construct a taxonomy, a list of all available variables (questions in the survey questionnaire) and options (values captured by the variable in the questionnaire) should be put together. Next, using a template such as the one here, each question and option on that list should be evaluated to determine which category they belong to for each of the dimensions of need, use case, product type and provider type. Most demand-side surveys will capture information pertaining to financial devices along two dimensions: products and providers. Therefore, these are listed in the taxonomy.

Question	Option	Need	Use case	Product type	Provider type	Variable	Option	Additional considerations

For coding purposes, the variable name and option (the value which the variable should be equal to) are also captured, along with additional considerations that need to be taken into consideration when creating the new needs variables.

This is an example of what the first three lines of a taxonomy could look like, based on the FinScope Zambia 2015 dataset<sup>3</sup>.

Question	Option	Need	Use case	Product type	Provider type	Variable	Option	Additional considerations
Where do you get most of the money from to buy/build the house? Question 1.8.2b	Chilimba	M	G	S	I	Q1_8_2B	9	
How/where will you get most of the money to pay for birth of a child (==1) (read option mentioned in Q3.5.1) If you have to? Question 3.5.2	Rely on savings group social fund	M	L	S	I	Q3_5_2	2	& Q3_5_1==1
	Sell something that I bought for this opurpose	M	L	S	K	Q3_5_2	12	& Q3_5_1==1

In the first row, the question relates to how the respondent could finance the house they own. The option in this instance (there may be other options captured elsewhere) was to use money from a Chilimba, a type of informal Zambian savings group. Therefore, the need that the respondent is meeting is to meet a goal (M) – by investing in a productive asset – to grow (G), using a savings (S) product from an informal (I) provider. The variable is called Q1\_8\_2B in the dataset and Chilimba, the option, is labelled number 9. No additional considerations need to be considered.

In rows two and three, respondents are indicating that they plan to manage the additional cost of a child by (1) relying on a savings group social fund or (2) selling something that they bought for his purpose. Both meet a goal (M) – to have a child – which is a type of life event (L), using savings (S). However, using savings from a savings group relies on an informal (I) provider, while buying an asset to sell later for a specific purpose constitutes savings in kind (K). Rows two and three illustrate two additional things. Firstly, both options have additional considerations attached to them: due to the manner in which the dataset was constructed, question Q3\_5\_1 needs to be equal to one in order for these lines to hold. Secondly, they show that a question can have multiple options. In fact, some questions have many options and each of them should be treated separately. Therefore, the list to which we referred in the beginning is a list containing all relevant variable-option pairs, each of which needs to occupy its own row in the taxonomy.

## Step 3: Coding and constructing an analysis of devices used towards needs

In the taxonomy, there are four dimensions along which variables and their options are categorised: needs, use cases, product types and provider types. Following construction of the taxonomy, as described in the box above, a binary variable is created for each option under each dimension. For example, whether each use case classifies under a specific need or not will be one binary variable. Whether it classifies under a specific product or provider type will create further binary variables. The number of binary variables that can be created depends on the number of options under each of the four dimensions. In other words, if there are four overarching needs, a total of 10 use cases, four product types and five provider types, a total of 23 binary variables need to be created.


Thereafter, these binary variables can be combined in numerous interesting ways. For example, the product types can be matched to need categories using these binary variables to explore what types of devices people are using towards each need.

One method that can be used to construct these binary variables, following on the discussion of the taxonomy, is to filter by letter for each of the dimensions (one dimension at a time) and to then export the resulting list of questions to data processing software, such as Stata, for coding purposes. For example, in the Figure 1, filtering the Need column and selecting M (“meeting goals”), without filtering Use case, Product type or Provider type, will result in a list in the Variable column which includes all the variables available to measure whether the “Meeting Goals” need is being met. This list can then be exported, along with the associated values in the Option column, to a software programme and then a binary variable can be constructed in whichever way is most convenient given the software selected.

Figure 1. Filtering example

Question	Option	Need	Use case	Product type	Provider type	Variable	Option	Additional considerations
Where do you get most of the money from to buy/build the house? Question 1.8.2b	Chilimba	M	G	S	I	Q1_8_2B	9	
How/where will you get most of the money to pay for birth of a child (==1) (read option mentioned in Q3.5.1) If you have to? Question 3.5.2	Rely on savings group social fund	M	L	S	I	Q3_5_2	2	& Q3_5_1==1
	Sell something that I bought for this opurpose	M	L	S	K	Q3_5_2	12	& Q3_5_1==1

3 Source: MAP Zambia publication authored by Cenfri and commissioned by Financial Sector Deepening Zambia. See: <http://cenfri.org/making-access-possible/map-zambia>



Thereafter, the Need column can be filtered by R (“Resilience”), leaving all the other columns unfiltered, then by L (“Liquidity”) and so on, until a binary variable has been created for all the categories in the Need column. Next, the Need column can be left unfiltered and the Use case column can be filtered, for each of its categories individually. The same can be done for the Product type and the Provider type columns.

The contents of the Options and Additional considerations columns will be taken into account depending on the software used. In Stata, for example, the contents of the Additional considerations column can be added as additional constraints while creating the variables.

## Step 4: Validate the approach

After constructing the binary variables that feed into constructs such as the needs strand, a validation of the results should be carried out to see if they are credible and stand strong alongside desktop research, qualitative demand-side findings or stakeholder consultations. Should there be a dramatic conflict with expectations, these may be explored in two ways:

- **Incorrect classification of question:** It is necessary to investigate whether the classification of the questions is correct. Some variables may initially appear to fit two needs categories. Often some reflection, along with contextual information, can provide clarification and reveal flaws in the categorisation. Some discretion will always be required.

- **Other issues with the dataset:** There are many potential stumbling blocks in survey design and data collection<sup>4</sup>. Therefore, results that deviate too strongly from expected outcomes can be investigated by returning to the data source. For example, checking whether certain questions could have been misinterpreted, either by the interviewer or the interviewee, whether the weight variable is correct, or whether skips were incorrectly included in the questionnaire, causing data to be collected on a sub-sample of the desired population. Where possible, the effect of these errors should be accounted for, but the levity of a particular situation can only be judged by the researcher.

**New insights.** If the questions are correctly classified and the data is a true depiction of reality, yet the analysis still yields results that differ from expectations, then the results may be able to reveal new insights in how people engage with financial services to serve their financial needs.

---

4 The literature on these topics is vast. A handbook such as the Handbook of Survey Design by Rossi et al. (2013) may serve as a good introduction to the field.

# 3 Transactional data analysis

## Step 1: Understand the nature and scope of the data

The point of departure in embarking on a transactional data analysis is to understand that transactional data analysis is significantly different from demand-side analysis. A large and important part of the initial analytical process is to explore the data to understand what it actually covers. It is important to look out for key differences to demand-side data, including:

- With demand-side data, a typical questionnaire format and design is followed, yielding a relatively standard cross-sectional dataset, which can answer a standard list of questions. Transactional data, on the other hand, is not standardised: it can comprise of any data captured by financial services providers, regulators, industry bodies and credit score companies, among others.
- The data usually already exists. Therefore, the researcher cannot influence or change the structure and contents, particularly when working with past data and not live data<sup>5</sup>.
- The data provider may not be transparent about the process through which the data was selected. Researchers therefore need to work with what they get from the provider and may have little view of the data collection process.
- Transactional data is likely to have a time component and not be strictly cross-sectional. However, in most instances, the type of data provided by financial service providers should not be treated as time-series data (as in macroeconomics). The time dimension is often too short to consider any real time-based effects. Therefore, it is typically best to treat transactional datasets as cross-sectional datasets, but, where relevant, to take time-related factors into account (such as seasonality or differences in the static picture from one point in time to another).
- Transactional data is normally not weighted, as you are working with the actual data rather than a sub-sample.
- Transactional datasets are likely to have far fewer columns or fields (tens) than the demand-side data and many more rows or transaction entries (hundreds of thousands and more).
  - The variables (columns) will likely include continuous responses, such as closing account balance, amount transacted, transaction fee, among others.
  - Non-continuous variables may have a few well-defined categories, in which case it is easy to tabulate these, investigate the different categories and check for irregularities. However, they may also contain many categories, for example, payment merchant codes can have more than a thousand categories. In this case the researcher should investigate whether the large number of categories is sensible or necessary and whether they can be bundled into a smaller, more manageable number of categories. Of course, any bundling methodology should be transparent and retain sufficient diversity to have explanatory power.

---

<sup>5</sup> It is unlikely that the provider will allow access to live data. Instead, data from the past year or another defined period may be provided.



## Step 2: Cleaning the data

**Removing blanks and outliers.** As with demand-side data, the first step in cleaning transactional data is to remove outliers (such as ages captured as zero), obviously incorrect entries and blanks in key fields.

### **Reduce computation memory requirements.**

Due to the large number of rows, data manipulations may require considerable processing time. Therefore, it is prudent to reduce the size of the dataset by changing variable types to make the process more efficient, without making any changes to the actual data<sup>6</sup>.

**Reduce the sample size.** It is not always necessary to work with the entire sample. A randomly drawn sub-sample may be sufficient to answer or test many questions<sup>7</sup>. Standard statistical techniques can be applied to determine the appropriate size of a sub-sample that is representative of the total sample to a specified certainty level.

## Step 3: Conducting the analysis


**Analytical framework.** The basic analytical framework will be the same as described for demand-side data analysis in Section 1. However, the nature of transactional data means that it is not suited to explore certain parts of the FinNeeds framework, such as financial needs and devices towards needs. The main elements of the transactional data analytical framework all centre around the usage element of the FinNeeds framework.

- **Usage intensity:** analysing transaction patterns to understand different usage dimensions, depending on the type of dataset
- **Usage segmentation:** segmenting users into groups or clusters to understand different user profiles
- **Usage determinants:** modelling usage to understand which demographic or other variables in the dataset best explains differences in usage behaviour

“ The main elements of the transactional data analytical framework all centre around the usage element of the FinNeeds framework. ”

<sup>6</sup> For example, certain computer programs capture data using variable lengths of characters (varchar). Others use a fixed character length (char), which would limit the storage size needed to process the data. The memory required by the researcher's software can often be reduced significantly by just changing the type of character used to capture information. In one insight2impact pilot study, the size of the dataset was reduced by changing string variables (words) to numeric variables (numbers) and then assigning labels to those numbers. This reduced computation times significantly.

<sup>7</sup> For example, some clustering algorithms have long computation times. However, these algorithms do not need to be applied to the entire sample to be accurate. Therefore, the researcher could use a representative sub-sample to create the clusters.



**General investigation.** As with survey data analysis, the first step is to do a general investigation of the data. This is arguably more important for transactional data than for demand-side analysis. If the data provider is not able to provide a data dictionary which summarises the data, the first step for the researcher is to create a basic data dictionary by using summary tools provided by the software package used.

### What is a data dictionary?

A data dictionary is a set of information describing the contents, format and structure of a database and the relationship between its elements. It is used to control access to and manipulation of the database.

A basic dictionary would include the following for each variable:

- Label
- Description: what it captures
- Type: char, varchar, etc.
- Number of missing observations
- For continuous variables: minimum, mean, median, maximum, standard deviation
- For discrete variables: number of categories
- Could also include: coefficient of variation, variance, skewness, kurtosis

The first step in the general investigation would be to use the data dictionary to identify the variables that are most useful to apply as part of the analytical framework. Then those variables can be used to create derivatives that fit with the analytical

framework. For example, insight2impact used the RFMD (recency, frequency, monetary value, duration) framework to understand usage in more depth:

- Recency and duration can both be calculated using “date of transaction” variables, which in themselves are not very useful (derivations of dates are often more useful, such as classifying users according to whether they have transacted in the past month or week).
- Frequency and monetary value are summaries of variables that capture individual transactions. For example, calculating the average number of transactions per week or month, or the average or total amounts transacted for a given period.

**Link transactions to use cases.** If the data allows, the usage-towards-needs part of the analytical framework can be explored by relating transactions to the use cases to which they apply. This can be done if transactions are assigned a code such as those used by VISA or Mastercard (so-called MCC codes), or internal codes generated by banks or payments switches. In such cases, transactions can be classified into different categories based on bundles of similar codes, for example, whether they are store purchases, petrol purchases, online purchase and, if so, on which online platform. Whether such an analysis is possible will depend on the type of dataset.

### Segmenting users via a clustering analysis.

As with the demand-side data analysis, it is useful to segment groups of customers in the transaction database, using clustering algorithms to create clusters of consumers based on the available variables. Typically, transactional data will have few demographic variables and often these are of



poor quality or outdated (demographic information is captured when the customer signs up and is never updated). The reasons why financial service providers capture customer data differs. Some capture to understand their clients better, others capture because it is a regulatory requirement. Either way, the information available is usually incomplete and can only yield a partial sketch of the customer. Therefore, including transactional data or derivatives such as those discussed above (RFMD) in the clustering exercise can lead to more informative clusters. Moreover, a clustering exercise that segments customers according to usage intensity serves the primary aim of the transaction analysis, namely to understand usage.

In the insight2impact pilot studies, an unsupervised k-means clustering methodology was used. Once the clusters have been generated, evaluate each cluster to establish what the characteristics of each cluster are (e.g. income category, education, gender and age). Clusters can then be named and compared according to their demographic profiles and transaction patterns (e.g. by indicating the percentage of each cluster that transacts daily, versus weekly, monthly or less frequently, or by showing the average number and value of transactions per month for each cluster).

**Model the determinants of usage.** The second key component of the analytical framework for transactional data analysis is to use the analysis to unpack the determinants of usage. This is done through a regression analysis where usage is set as the dependent variable and the statistical significance of different variables in the dataset, notably demographic variables, is then tested in explaining usage. In the insight2impact pilot studies, an ordered logistic regression analysis was applied. Usage intensity was defined according to the RFMD framework and the statistical significance and strength of different demographic variables was then tested through the modelling exercise.

“ Some [financial service providers] capture [customer data] to understand their clients better, others capture because it is a regulatory requirement. ”

# 4 Merged demand and transactional data analysis

## Step 1: Understand what merged data is and can offer

**Why merged data?** As explained in the FinNeeds toolkit, there are benefits to exploring FinNeeds from both a demand-side data and transactional data perspective. While survey data is rich in demographic details<sup>8</sup>, gives insights into respondents' full financial lives, including their use of informal devices, and can relate device usage towards the use cases and financial needs underlying the uptake of financial services, it cannot portray actual engagement with financial services in a detailed or accurate way. Respondents may not be able to recall their exact transaction profiles, or may be reluctant to disclose the details. Transactional data, on the other hand, captures actual transaction frequency and amounts. Thus, it complements survey data by providing a granular and objective picture of usage patterns. However, transactional data cannot be representative of a whole target population. Neither can it explain behaviour, including use cases, device portfolios or reasons for device choice, outside of the interactions of the client with the financial institution in question. To generate rich demographic and use case and usage choice information, while eliminating the reliance on self-reporting of transaction profiles, it is necessary to create a merged demand and transactional dataset.

**What is a merged dataset?** A merged dataset is generated when a demand-side survey is administered to individuals in a transactional dataset. The responses are then matched via a unique identifier to the transaction entries (rows) for the person in question in the transactional dataset. Thus, for the purpose of applying the FinNeeds framework, a merged dataset is a combination of the demand-side and transactional data sources discussed, where the demand-side information can be linked to the transactional (supply-side) information for the same individual. In doing so, however, one reverts back to a smaller sample of data. The downside of merged data is that it cannot render data analysis that is representative or form the basis for statistical modelling and regressions.

## Step 2: Cleaning the data


**Standard data cleaning techniques.** The demand-side component of the merged dataset should be cleaned as described for the demand-side analysis, while the transaction component should be cleaned as described for the transaction component.

**Consistency checks.** An additional requirement for a merged dataset is to conduct checks to ensure that the respondents from the demand-side model are merged with the correct individuals in the transactional dataset. Other than relying on the financial service provider to provide the correct information for the merge, the researcher can also check the demographic information in the demand-side data against that in the transactional dataset, allowing for time-related changes where variables are available in both datasets. For example, the

---

<sup>8</sup> This provides additional value to the financial service provider, as many providers have a poor understanding of their consumer base due to the difficulty of collecting accurate information on them. For example, a bank could improve its credit-scoring methodology by understanding more about the consumers who tend to default (from their demand-side information).





researcher could verify that the age, gender and location of respondents in the demand-side and transactional datasets are the same (or at least very similar). It is unlikely that these characteristics will be consistently similar by chance.

## Step 3: Conducting the analysis

### **Standard FinNeeds analytical framework**

**and approach.** The same FinNeeds analytical framework as set out for survey data analysis in Section 1 applies to the merged dataset. Likewise, similar methods are advised for analysing merged and demand-side data. However, the additional depth of the data can improve the value of these methods in a merged dataset:

- The transactional data component can be used to verify whether the amounts reported in demand-side data are accurate or not, and allow researchers to construct confidence intervals around the accuracy with which usage data is reported in demand-side data.
- The demand-side data component can lead to more useful clusters of individuals and to understand use case incidence and device uptake beyond what is witnessed in the transactional data.

### **Enhancing the determinants of usage model.**

Where the merged dataset sample size is large enough to enable regression analysis, the merged data can be used to enhance the determinants of the usage modelling exercise. This is done by adding additional demographic variables as well as information on use case incidence to the usage model and then testing the statistical significance of such variables compared to those in the transactional dataset.

### **Providing a device portfolio overview to put**

**usage patterns in context.** A merged dataset can also be used to profile the device portfolio mix (such as formal versus informal, cash versus digital) for respondents in the transaction database. For example, when distinct usage clusters are created based on the transactional data, the merged dataset can then describe the device portfolio, including informal, social and personal devices, taken up by each usage cluster towards each need category. For example, a certain cluster may be shown to transact relatively frequently on their bank accounts, but when looking at the merged data it may show that most of their needs are actually still met outside of the banking system.

---

31 JUMO. (2018). Uber partners with JUMO to provide driver partners with vehicle finance.



# How to find us:

Get involved. Contact us.

+27 21 913 9510  
[i2ifacility.org](http://i2ifacility.org)

Join the conversation:  
**#FinNeeds**

-  @i2ifacility
-  /insight2impact
-  /insight2impact
-  /i2ifacility

Established and powered by



Sponsored by

BILL & MELINDA  
GATES *foundation*

