

サブセット割合の理解と利用について

キー項目のサブセット割合を調べることは、データの質を保証するのに必要な作業です。

サブセット割合は、データモデルビューア内でキー項目を選択するとプレビューに表示されます。

以下は「受注明細」「顧客マスタ」のデータモデルですが、「顧客ID」項目をクリックすると、「サブセット割合」「ユニーク値総数」「現在のユニーク値の数」が確認できます。

▼プレビュー

軸として追加	顧客ID [パーフェクトキー]	顧客マスタ
メジャーとして追加	密度 100%	顧客ID 顧客名 地域 県 年齢 性別 年代
	サブセット割合 100%	1 平野 祐子 中部 愛知 23 女性 20代
	複製を含む false	2 松井 美恵子 関西 大阪 30 女性 30代
	ユニーク値総数 2020	3 白石 真美子 関西 和歌山 34 女性 30代
	現在のユニーク値の数 2020	4 花岡 美紀子 関西 兵庫 20 女性 20代
	非NULL値の数 2020	5 池内 麻美 関西 兵庫 21 女性 20代
	タグ Snumeric Sinteger Skey	6 堀江 里香 関西 兵庫 22 女性 20代
		7 富沢 智美 関西 兵庫 18 女性 10代
		8 星野 文恵 関西 兵庫 22 女性 20代

<各項目の説明>

サブセット割合	選択している項目のユニーク値の合計数に対して、 テーブル内で選択した項目のユニーク値の数の割合を表示します。
ユニーク値総数	モデル内の全てのテーブル「受注明細」「顧客マスタ」の中から、 選択している項目「顧客ID」の全てのユニーク値をカウントします。
現在のユニーク値の数	現在選択したテーブル「顧客マスタ」の中から、 選択している項目「顧客ID」内のユニーク値をカウントします。

全頁の「顧客マスタ」は、「顧客ID」項目の100%が「受注明細」に表示されています。
これは、100%のサブセット割合を「顧客マスタ」から見つけることができます。

次に「受注明細」のサブセット割合も確認してみます。



▼プレビュー

軸として追加	顧客ID	受注明細
メジャーとして追加	密度 100%	顧客ID 商品コード 受注ID 数量 特価ID 販売価格
	サブセット割合 74.2%	1 707 45038 2 1 3500
	複製を含む false	2 708 45038 1 1 3500
	ユニーク値総数 2020	3 709 45038 2 1 950
	現在のユニーク値の数 1500	4 712 45038 3 1 890
	非NULL値の数 1500	5 714 45038 1 1 4900
	タグ Snumeric Sinteger Skey	6 715 45038 4 1 4900
		7 741 45038 2 1 136000
		8 742 45038 1 1 136000

このテーブルから、サブセット割合が100パーセントよりも少ない事がわかります。

2020(=74.2%)のユニーク顧客のみが「受注明細」の中に表示されていますが、

100パーセントを下回るサブセット割合は「受注明細」のようなデータでは普通の状態です。

もし、注文のない顧客データを含めたくなければ、“Where Exists(顧客ID)”句をロードスクリプトに挿入すれば、「顧客マスタ」から不要な顧客を削除できます。

全頁までは100%となるサブセット割合をみてきましたが、「顧客マスタ」の値が100パーセントを下回った場合を確認してみます。

受注明細

顧客ID
商品コード
受注ID

顧客マスタ

顧客ID
顧客名
地域

▼プレビュー

軸として追加

メジャーとして追加

顧客ID	密度	100%
サブセット割合	99.7%	
複製を含む	false	
ユニーク値総数	2005	
現在のユニーク値の数	2000	
非NULL値の数	2000	
タグ	Snumeric Sinteger Skey	

顧客マスタ

顧客ID	顧客名	地域	県	年齢	性別	年代
1	平野 祐子	中部	愛知	23	女性	20代
2	松井 美恵子	関西	大阪	30	女性	30代
3	白石 真美子	関西	和歌山	34	女性	30代
4	花岡 美紀子	関西	兵庫	20	女性	20代
5	池内 麻美	関西	兵庫	21	女性	20代
6	堀江 里香	関西	兵庫	22	女性	20代
7	富沢 智美	関西	兵庫	18	女性	10代
8	星野 文恵	関西	兵庫	22	女性	20代

上記では、5つの顧客ID（2005→2000）を見落としているということことがわかります。

これは、「顧客マスタ」には存在するが、「受注明細」に存在していないデータロードしてしまった事が原因で件数に差異が発生しています。

見落とされた顧客IDはシート上で、2つの列を持つテーブルを作成すれば確認できます。

顧客名がnull値 (-) で表示されたものが該当データであることがわかります。

顧客ID	顧客名
1998	佐々木一郎
1999	野島 雪枝
2000	上松 真智子
11501	-
11502	-
11503	-
11504	-
11505	-

以上のように、データモデル作成時にサブセット割合を調べることは重要なステップとなります。

また、データモデルの質を高めることで不要なデータ確認等の作業を削減できます。