

Aquisição e Preparação de Dados

Parte 2

Cláudio Roberto de Lima Martins

Msc Computação / Prof. IFPA

claudiomartins2000@gmail.com



Agenda

- Estatísticas descritivas
- AMC
- Ferramentas
 - SQL
 - Open Refine
 - Saiku

Estatísticas Descritivas

Estatística Descritiva

A estatística descritiva é empregada na análise de dados para descrever e resumir os dados.

A Estatística Descritiva permite-nos resumir, descrever e compreender os dados de uma distribuição/conjunto, é organizada assim:

Análise univariada (uma variável):

- A) **medidas de tendência central** (média, mediana e moda),
- B) **medidas de dispersão** (valores mínimo e máximo, desvio padrão e variância), **percentis, quartis e decis**, e
- C) **medidas de distribuição** (achatamento e simetria da curva de distribuição).

Análise bivariada (mais de uma variável): analisa o relacionamento entre duas variáveis diferentes.

Medidas quantitativas de dependência incluem [correlação](#) (como o [coeficiente de correlação de Pearson](#) quando ambas variáveis são contínuas, ou [Coeficiente de correlação de postos de Spearman](#) quando ambas variáveis são discretas) e [covariância](#).

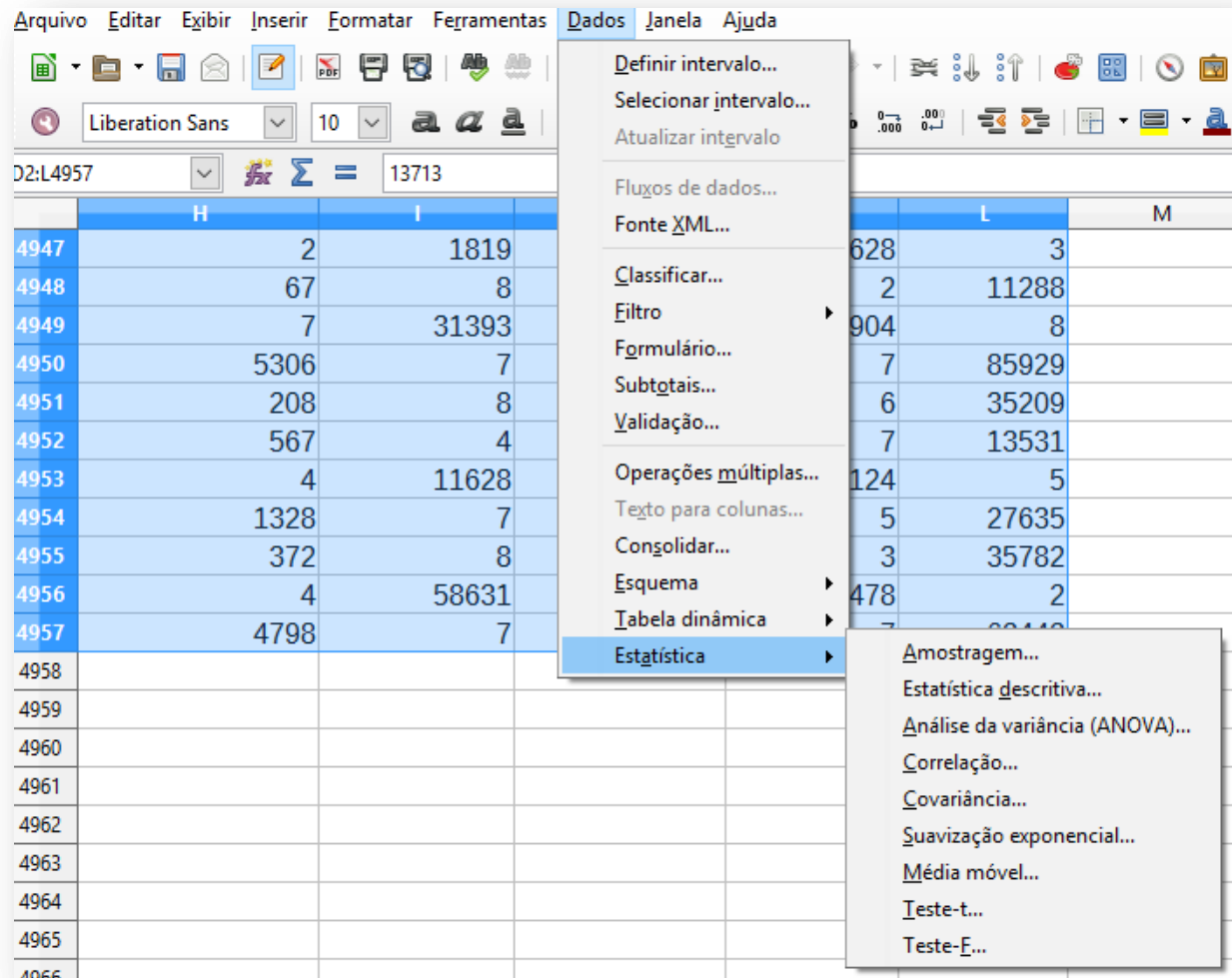
Funções no Calc

- Supondo que os dados de uma coluna estejam no intervalo A1 a A30 pode-se utilizar as seguintes funções:

Função	Fórmula
Contagem Numérica	=CONT.NÚM (A1 : A30)
Mínimo	=MÍNIMO (A1 : A30)
Máximo	=MÁXIMO (A1 : A30)
Total (Soma)	=SOMA (A1 : A30)
Média	=MÉDIA (A1 : A30)
Moda	=MODO (A1 : A30)
Mediana	=MED (A1 : A30)
Variância	=VAR (A1 : A30)
Desvio padrão	=DESVPAD (A1 : A30)

Assistente de Estatística do Calc

- Fornece forma rápida de gerar funções e operações estatísticas no Calc.
- Para isso, selecione o intervalo de dados (numéricos) e escolha: Dados > Estatística



Ferramentas

**(Calc, SQL, Open Refine,
Saiku)**

Calc e Excel

- Ferramenta para uso de Planilhas Eletrônicas,
 - Calc é um produto do Libre Office, um software livre.
 - Excel possui recursos avançados para análise de dados (instalados como plugins)

Mais em : <http://www.cultura.ufpa.br/dicas/open/calc-ind.htm> e <http://www.es.iff.edu.br/softmat/projetotic/portaltic/projetotic/download/atividades1/ApostilaCalc.pdf>

SQL (Linguagem de Consulta em BD)

- Linguagem usada na manipulação dos dados armazenados em forma de tabelas de um gerenciador de banco de dados (SGBD).
- No SGBD Postgresql (versões 8 ou maior), o SQL segue o padrão ANSI 2008.
- Existem dois tipos de comandos:
 - DDL que definem e criam objetos no banco
 - DML que manipulam os dados (consultas e atualização)
- A seguir veremos um resumo dos principais comandos usados para obter e tratar dados.

Ver documentação de referência em português para o Postgres 8.2:

<http://pgdocptbr.sourceforge.net/pg82/reference.html>

Comando CREATE TABLE

- **CREATE TABLE** -- cria uma tabela.
- Ex: considere a tabela para representar os dados extraídos do IBGE para o Censo Agro 1995 da atividade anterior:

```
CREATE TABLE agro1995 (  
    munic                                VARCHAR,  
    nome_municipio                      VARCHAR,  
    sigla_uf                            VARCHAR,  
    lavoura_perm                        VARCHAR,  
    lavoura_temp                        VARCHAR,  
    lavoura_temp_desc                  VARCHAR,  
    pastagens_naturais                  VARCHAR,  
    pastagens_plant                     VARCHAR,  
    matas_f_nat                         VARCHAR,  
    matas_flor_art                     VARCHAR,  
    terras_prod                         VARCHAR,  
    terras_inap                         VARCHAR  
);
```

Comando COPY

- **COPY** -- copia dados entre um arquivo e uma tabela. Serve para importar e exportar dados.
- O comando **COPY FROM** copia dados *de* um arquivo para uma tabela (adicionando os dados aos já existentes na tabela).

Ex: COPY AGRO1995 FROM
 'C:/TMP/agro-munic-1995-utf8.csv'
 CSV HEADER DELIMITER ';' ;

- O comando **COPY TO** copia o conteúdo de uma tabela *para* um arquivo.
 - O comando **COPY TO** também pode copiar os resultados de um comando SELECT.

Ex:
COPY (SELECT * FROM AGRO1995)
 TO 'C:/TMP/agro1995-utf8.csv'
 CSV HEADER DELIMITER '|' ;

Comando SELECT

- SELECT -- retorna linhas de uma tabela ou mais tabela (incluindo dados de visão).
- Ex: SELECT * FROM agro1995;

Atividades

- Crie uma tabela “TERRITORIO” para armazenar os municípios e seus mapeamentos AMC, com a seguinte estrutura:

ano_ref	VARCHAR (4)
ano_amc	VARCHAR (4)
populacao	INTEGER
codigo	VARCHAR (10)
nome	VARCHAR
mun_cod_amc	VARCHAR
nome_amc	VARCHAR
amazonia	VARCHAR (1)
latitude_amc	VARCHAR
longitude_amc	VARCHAR
amc_original	VARCHAR
tipo_territorio_id	INTEGER

ATIVIDADE

- Carregue com o comando COPY os dados dos arquivos territorio2006-1995.csv, territorio2000-2000.csv e territorio2010-2000.csv para a tabela TERRITORIO.
- Elabore uma consulta para retornar os municípios de 2000 e suas populações.
- Elabore uma consulta para retornar os municípios de 2010 e suas populações.
- Elabore uma consulta para retornar os municípios de 2010 agregados em 2000 (AMC) e as populações.
- Elabore uma consulta que mostre os municípios em AMC com as populações em 2000 e 2010.

Tabela AMC (Áreas Mínimas Comparáveis)

- Um dos problemas mais comuns para os pesquisadores que trabalham com dados municipais, com informações para diferentes anos, é o problema de compatibilização das malhas de municípios (e comparação de dados).
- Entre 1990 e 2010, quase mil e quinhentos municípios novos foram criados.
 - Em uma análise sobre os dados geográficos, o pesquisador, ao comparar dados de anos diferentes, pode encontrar, por exemplo, que a população de um determinado município está diminuindo, simplesmente por que o município original foi dividido em dois novos, sendo que um deles manteve o mesmo nome (e mesmo código) do município de origem.

Tabela AMC (Áreas Mínimas Comparáveis)

- Para minimizar esse problema, o Ipea e o IBGE desenvolveram o conceito de **Áreas Mínimas Comparáveis – AMC**.
- As áreas mínimas comparáveis são uniões de um ou mais municípios, de forma que o analista tenha áreas mínimas de referências, que sejam imutáveis ao longo dos anos.
- Com isso, o analista poderá agregar informações, em um determinado ano, para os municípios de cada AMC.
- Como as AMC são constantes ao longo dos anos, todas as análises de comparação do dinamismo dos municípios, por exemplo, podem ser feitos ao nível de AMC's.

Censo com AMC (1872-2000) .

Brasil: Número de municípios nos Censos e AMC nos períodos intercensitários, 1872-2000			
Anos censitários	Número de municípios	Período intercensitário	Número de AMC
1872	643	1872-2000	432
1920	1305	1920-2000	952
1940	1575	1940-2000	1275
1950	1891	1950-2000	n.d.
1960	2768	1960-2000	2407
1970	3974	1970-2000	3659
1980	3991	1980-2000	3692
1991	4491	1991-2000	4267
2000	5507	-	-

Fonte: IBGE e IPEA.

Obs.: O Censo de 2000 utilizou a malha municipal de 1997.

Tabelas AMC (Ipea)

O Ipea fornece os seguintes mapeamentos, com os dados de população do Ano inicial.


ANO		AMC ORIGEM
2000	→	1997
2001	→	1997
2002	→	1997
2003	→	1997
2004	→	1997
2005	→	1997
2006	→	1997
2007	→	1997
2008	→	1997
2009	→	1997
1970	→	1970
1980	→	1970
1991	→	1970
1991	→	1991
1993	→	1970
1993	→	1991
1997	→	1970
1997	→	1991
1997	→	1997
1998	→	1997
1999	→	1997

Atividade


- Crie uma base de dados e o esquema para representar o mapeamento AMC para os anos 2010 a 1970 (conforme visto na tabela anterior).
- Simule um caso de consulta à AMC de 2005 para 1997 de uma base de dados de entrada (por exemplo, Credito Rural).

- **OpenRefine** (antes, *Google Refine*) é uma ferramenta para trabalhar dados desorganizados.
- Serve para limpeza e transformação de dados
- Pode trabalhar com dados obtidos em serviços web e arquivos externos e locais (na mesma máquina onde está o aplicativo).
- Roda como aplicação-web local (servidor Tomcat)

OpenRefine (tela inicial)

 *A power tool for working with messy data.*

Create Project
Open Project
Import Project


Version 2.5 [r2407]

[Help](#)
[About](#)

Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with Google Refine extensions.

Get data from

This Computer
[Web Addresses \(URLs\)](#)
[Clipboard](#)
[Google Data](#)

Locate one or more files on your computer to upload:

Procurar...

Next »

Ferramenta: OpenRefine

Marine Facilities - Google Refine

127.0.0.1:3333/project?project=1553048151762

Google refine Marine Facilities Permalink

Open... Export Help

Facet / Filter Undo / Redo

400 rows

Show as: rows records Show: 5 10 25 50 rows

Extensions: Freebase

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? Watch these screencasts

1. 27.85170173 -82.82377167 83rd Ave. Bridge 83rd ave. n. and talahassee dr. St Pete Pinellas 33708 813-393-2090 O Other

2. 27.80251502 -82.80054474 ABC Marina 206 150th avenue Madeira Beach Pinellas 33708 813-393-2090 M Marina

3. 28.17527007 -82.78726959 Anclote Gulf Park, Fishing Pt. on seminole, off anclote Tarp Springs Pinellas O Other

4. 28.162714 -82.7693634 Anclote Mart Bay

5. 27.75236128 -82.74485016 Bay

6. 27.74250029 -82.74588776 Bay

7. 27.81745909 -82.76763916 Bay

8. 27.70397567 -82.63999176 Bay

9. 27.70954131 -82.72505951 Bay Bridge

10. 27.91724777 -82.6264389 Belle Case

11. 27.94431876 -82.81217957 Belle Rose

75129 rows 1st Zone

Show as: rows records Show: 5 10 25 50 rows 2nd Zone

All university 3rd Zone endowment numFaculty

1. Facet 15 5500

2. Text filter 15 5500

3. Edit cells y Lyon 2 121

4. Edit column 4700000

5. Transpose 16586100

6. Sort... 16586100

7. View 40200750 838

8. Reconcile 40200750 838

9. 40200750 838

10. 40200750 838

11. Idaho State University 40200750 838

12. Idaho State University 40200750 838

13. Idaho State University 40200750 838

22

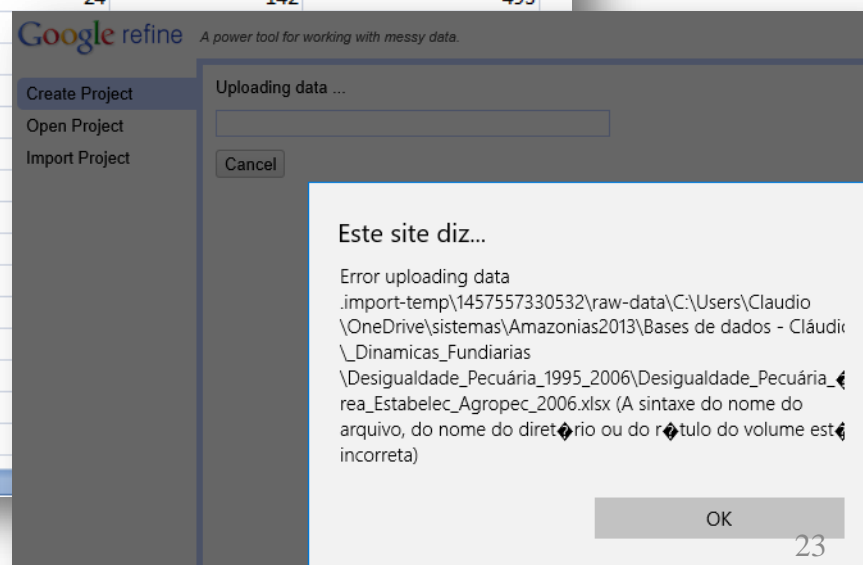
Exemplo (no OpenRefine):

Desigualdade_Pecuária_Área_Estabelec_Agropec_2006.xlsx

- Crie um projeto, tentando abrir o arquivo

Desigualdade_Pecuária_Área_Estabelec_Agropec_2006.xlsx

- Como o formato da planilha está fora do padrão, ocorre um erro.
- Desta forma, ou você elimina as linhas que provocam o erro ou copia o intervalo de dados (incluindo o nome das colunas). Em seguida, tente Criar o projeto colando os dados em “Clipboard”



Colando dados (Clipboard)

Google refine

A power tool for working with messy data.

Create Project

Open Project

Import Project

Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with Google Refine extensions.

Get data from

This Computer

Web Addresses
(URLs)

Clipboard

Google Data

Paste data from clipboard here:

CODIGO	UF	Município 2005-2008 em ordem de UF e nome					Menos de 1 há		
		De 1 a menos de 2 ha	2 a menos de 5 ha	De 5 a menos de 10 ha	De 10 a menos de 20 ha	De 20 a menos de 50 ha	De 50 a menos de 100 ha	De 100 a menos de 200 ha	De 200 a menos de 500 ha
		De 100 a menos de 200 ha	De 200 a menos de 500 ha	De 500 a menos de 1000 ha	De 1000 a menos de 2500 ha	De 2500 ha e mais			
5200050	GO	ABADIA DE GOIAS	0	X	30	106	255	963	
		1.235	1.841	3.511	X	-	-		
3100104	MG	ABADIA DOS DOURADOS	8	24	142	493	2.082		
		7.623	11.428	17.449	22.751	6.697	X	-	
5200100	GO	ABADIANIA	11	36	500	1.415	2.968	10.459	
		10.888	12.190	18.743	13.351	9.978	X		
3100203	MG	ABAETE	1	9	102	264	1.343	7.712	15.359
		24.100	38.144	27.007	11.788	12.571			
1500107	PA	ABAETETUBA	386	939	3.071	4.900	6.378	13.918	
		13.905	14.076	18.960	3.940	X	X		
2300101	CE	ABAIARA	151	269	484	512	718	1.580	1.315
		966	1.679	X	-	-			
2900108	BA	ABAIRA	32	377	2.116	1.598	1.945	3.217	1.813
		869	1.589	-	-	-			
2900207	BA	ABARE	78	158	674	312	1.169	5.182	3.564
		2.687	5.847	X	X	X			
4100103	PR	ABATIA	20	22	632	1.614	2.672	3.754	2.591
		2.991	2.880	X	-	-			
4200051	SC	ABDON BATISTA	3	13	290	885	2.580	6.448	
		3.558	1.364	1.928	X	-	X		
1500131	PA	ABEL FIGUEIREDO	0	-	0	X	105	1.849	
		2.923	4.370	15.866	6.286	12.739	10.261		

Next »



Version 2.5 [r2407]

Criando o projeto

A power tool for working with messy data.

[« Start Over](#) [Configure Parsing Options](#) Project name [Create Project »](#)

	CODIGO	UF	Município 2005-2008 em ordem de UF e nome	Menos de 1 há	De 1 a menos de 2 ha	2 a menos de 5 há	De 5 a menos d
1.	5200050	GO	ABADIA DE GOIAS	0	X	30	
2.	3100104	MG	ABADIA DOS DOURADOS	8	24	142	
3.	5200100	GO	ABADIANIA	11	36	500	
4.	3100203	MG	ABAEETE	1	9	102	
5.	1500107	PA	ABAEETETUBA	386	939	3.071	
6.	2300101	CE	ABAIARA	151	269	484	
7.	2900108	BA	ABAIRA	32	377	2.116	

CSV / TSV / separator-based files
[Line-based text files](#)
[Fixed-width field text files](#)
[PC-Axis text files](#)
[JSON files](#)
[RDF/N3 files](#)
[XML files](#)
[Open Document Format spreadsheets \(.ods\)](#)
[RDF/XML files](#)
[Excel \(.xlsx\) files](#)
[Excel files](#)

Columns are separated by

☐ commas (CSV)
☒ tabs (TSV)
☐ custom

Escape special characters with \

☐ Ignore first 0 line(s) at beginning of file
☒ Parse next 1 line(s) as column headers
☐ Discard initial 0 row(s) of data
☐ Load at most 0 row(s) of data
☒ Parse cell text into numbers, dates, ...
☒ Quotation marks are used to enclose cells containing column separators

☒ Store blank rows
☒ Store blank cells as nulls
☐ Store file source (file names, URLs) in each row

Visualização dos dados no OpenRefine

Google refine Pecuária_Área_Estabelec_Agropec_2006 Permalink

Open... Export ▾ Help

Facet / Filter Undo / Redo 0

5564 rows Extensions: Freebase ▾

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

Using facets and filters 

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

▼ All	▼ CODIGO	▼ UF	▼ Município 2005-	▼ Menos de 1 há	▼ De 1 a menos de	▼ 2 a menos de 5	▼ De 5 a menos de
☆ ↗ 1.	5200050	GO	ABADIA DE GOIAS	0	X	30	106
☆ ↗ 2.	3100104	MG	ABADIA DOS DOURADOS	8	24	142	493
☆ ↗ 3.	5200100	GO	ABADIANIA	11	36	500	1.415
☆ ↗ 4.	3100203	MG	ABAETE	1	9	102	264
☆ ↗ 5.	1500107	PA	ABAETETUBA	386	939	3.071	4.9
☆ ↗ 6.	2300101	CE	ABAIARA	151	269	484	512
☆ ↗ 7.	2900108	BA	ABAIRA	32	377	2.116	1.598
☆ ↗ 8.	2900207	BA <small>edit</small>	ABARE	78	158	674	312
☆ ↗ 9.	4100103	PR	ABATIA	20	22	632	1.614
☆ ↗ 10.	4200051	SC	ABDON BATISTA	3	13	290	885

Tutorial OpenRefine:

http://enipedia.tudelft.nl/wiki/OpenRefine_Tutorial

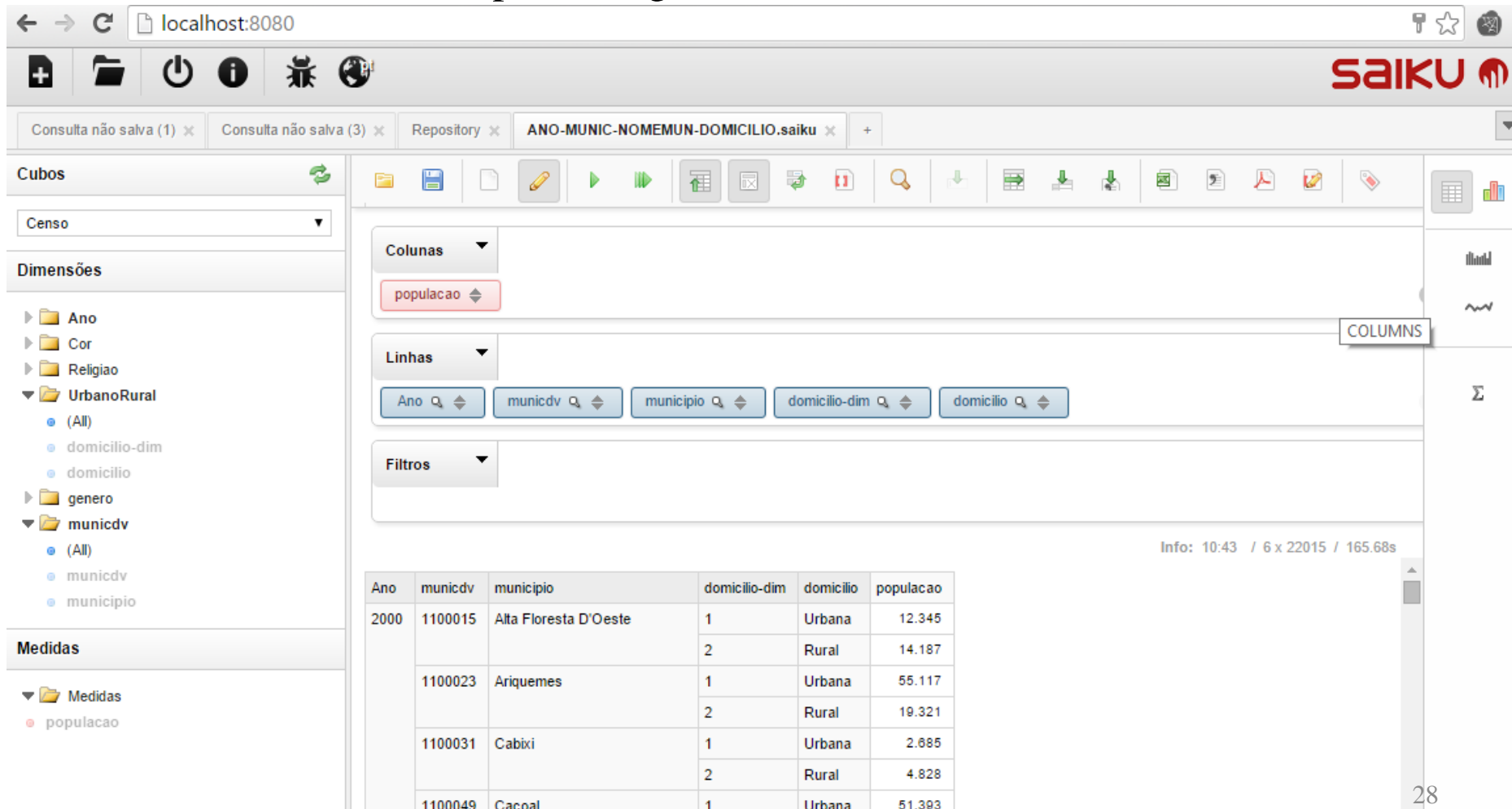
Ferramenta “Saiku”



- Saiku é uma ferramenta OLAP, permitindo que os usuários explorem dados carregados a partir de um banco de dados multidimensional.
- Emprega uma interface Web que facilita o uso do “arrastar e soltar”, tudo dentro de um navegador (browser).
- O usuário pode selecionar os dados do seu interesse, permitindo a exploração a partir de diferentes perspectivas e detalhes.
- A ferramenta permite salvar os resultados (consultas), compartilhá-las, exportá-las para Excel e CSV, gerar estatísticas básicas e gráficos em diferentes formatos.

Exemplo de construção de uma consulta no Saiku

A título de demonstração, foram carregados os dados do censo 2000 e 2010, com variáveis básicas para servir de estudo de caso. Esses dados estão no cubo “Censo”. No exemplo, para os censos de 2000 e 2010, recupera-se os valores da população para as dimensões: Ano, Município (codigo e nome) e Domicilio (rural e urbano)



The screenshot displays the Saiku OLAP tool interface. The browser address bar shows 'localhost:8080'. The interface includes a sidebar with dimensions (Cubos, Dimensões) and measures (Medidas). The main area shows the query configuration for the 'Censo' cube. The columns section contains 'populacao'. The lines section contains 'Ano', 'municdv', 'municipio', 'domicilio-dim', and 'domicilio'. The filters section is empty. The results table shows population data for the year 2000 across different municipalities and domiciles.

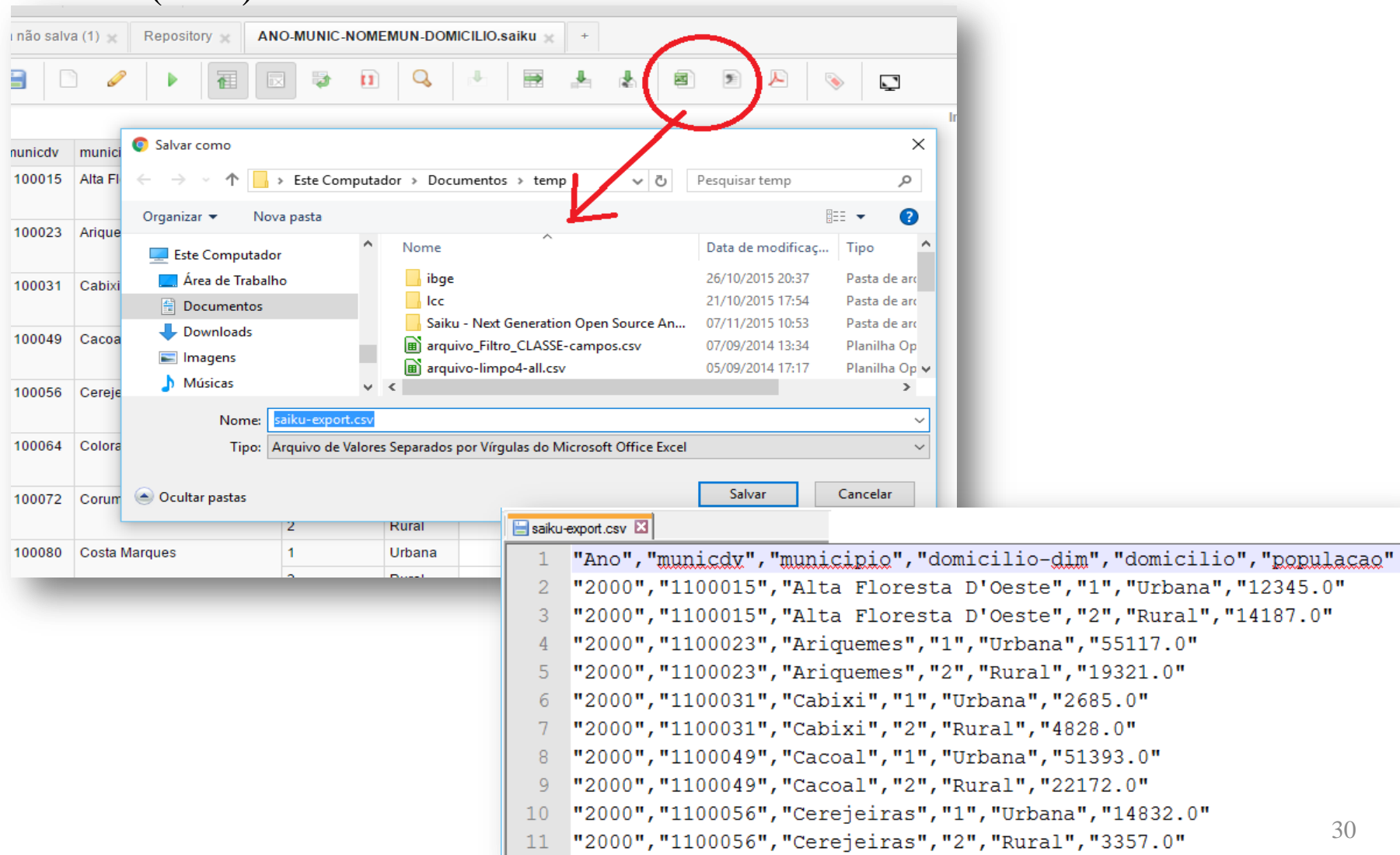
Ano	municdv	municipio	domicilio-dim	domicilio	populacao
2000	1100015	Alta Floresta D'Oeste	1	Urbana	12.345
			2	Rural	14.187
	1100023	Ariquemes	1	Urbana	55.117
			2	Rural	19.321
	1100031	Cabixi	1	Urbana	2.685
			2	Rural	4.828
	1100049	Cacoal	1	Urbana	51.393

Esquema do cubo (arquivo SchemaCenso.xml)

```
<?xml version="1.0"?>
<Schema measuresCaption="populacao" name="censo">
  - <Cube name="Censo" enabled="true" cache="true" visible="true" caption="Censo">
    <Table name="censopessoa" alias="censopessoa" schema="cubo"> </Table>
    - <Dimension name="Ano" visible="true" highCardinality="false" type="StandardDimension">
      - <Hierarchy visible="true" hasAll="true">
        <Level name="Ano" visible="true" type="Integer" hideMemberIf="Never" levelType="Reg
        </Hierarchy>
      </Dimension>
    - <Dimension name="municdv" foreignKey="municdv">
      - <Hierarchy hasAll="true" primaryKey="muncoddv">
        <Table name="munic_geo2000_2010_tratados" schema="territorio"/>
        <Level name="municdv" type="String" uniqueMembers="true" column="muncoddv"/>
        <Level name="municipio" column="munnome"/>
        <Level name="amazonia" column="amazonia"/>
      </Hierarchy>
    </Dimension>
    - <Dimension name="cor-dim" caption="Cor" foreignKey="cor">
      + <Hierarchy hasAll="true" primaryKey="cor_cod">
    </Dimension>
    - <Dimension name="UrbanoRural" foreignKey="domicilio">
      - <Hierarchy hasAll="true" primaryKey="domicilio_cod">
        <Table name="domicilio" schema="cubo"/>
        <Level name="domicilio-dim" type="String" column="domicilio_cod"/>
        <Level name="domicilio" type="String" column="domicilio_nome"/>
      </Hierarchy>
    </Dimension>
    - <Dimension name="Religiao">
      - <Hierarchy hasAll="true">
        <Level name="religiao-dim" type="String" column="religiao"/>
      </Hierarchy>
    </Dimension>
    - <Dimension name="genero" foreignKey="genero">
      - <Hierarchy hasAll="true" primaryKey="id">
        + <InlineTable alias="sexo">
          <Level name="genero-dim" uniqueMembers="true" column="id" nameColumn="desc"/>
        </Hierarchy>
```

Exportação do dados da consulta (no Saiku)

A ferramenta permite exportar os dados da consulta em dois formatos: CSV e Planilha (XLS).



The screenshot illustrates the export process in Saiku. The main window shows a table with columns 'municdv', 'munic', and 'domicilio'. A red circle highlights the export icon in the toolbar, and a red arrow points to the 'Salvar como' dialog box. The dialog shows the file 'saiku-export.csv' being saved in the 'Documents' folder. Below the dialog, a preview of the CSV data is shown.

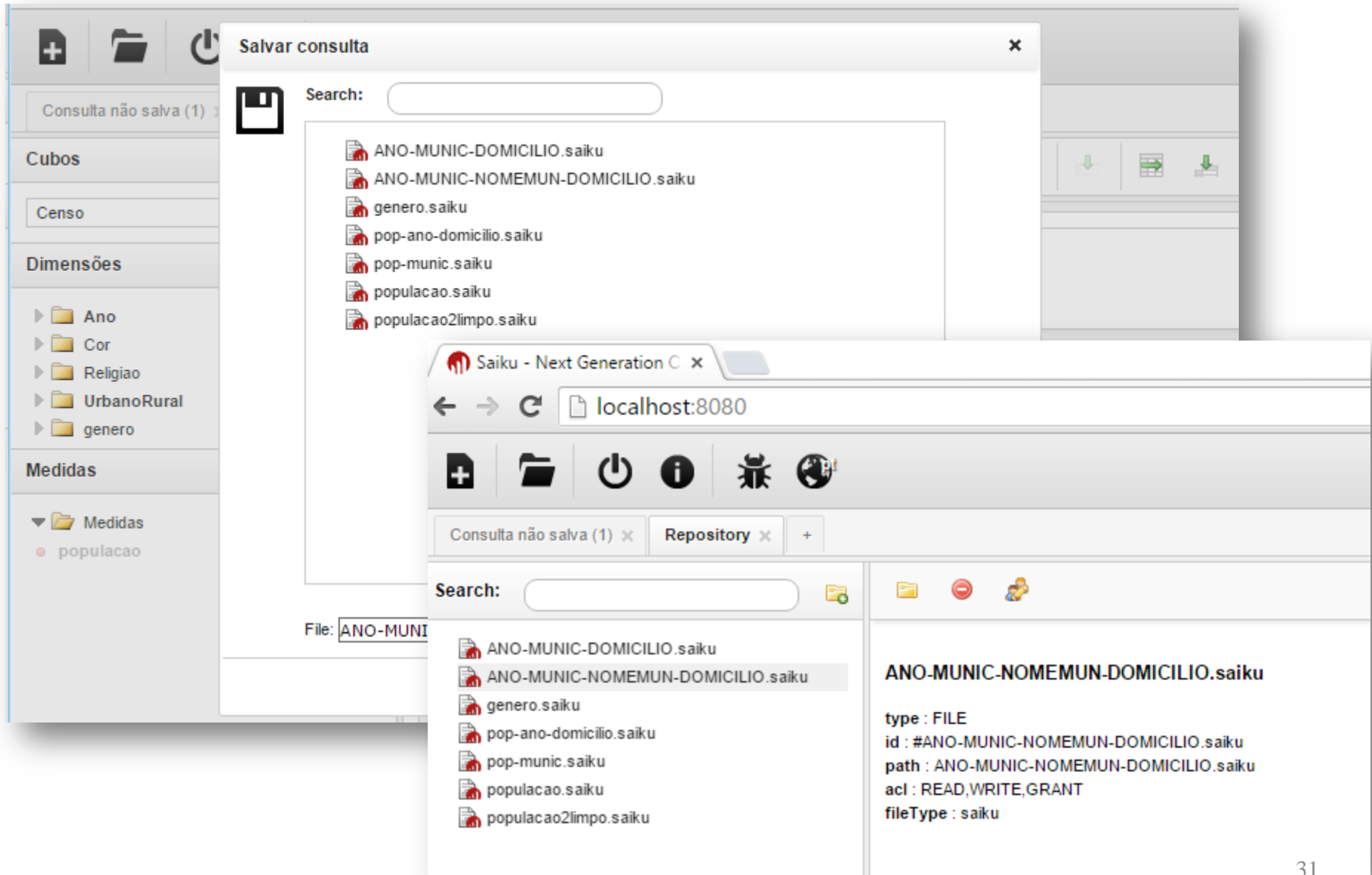
Nome: saiku-export.csv
Tipo: Arquivo de Valores Separados por Vírgulas do Microsoft Office Excel

saiku-export.csv

```
1 "Ano","municdv","municipio","domicilio-dim","domicilio","populacao"
2 "2000","1100015","Alta Floresta D'Oeste","1","Urbana","12345.0"
3 "2000","1100015","Alta Floresta D'Oeste","2","Rural","14187.0"
4 "2000","1100023","Ariquemes","1","Urbana","55117.0"
5 "2000","1100023","Ariquemes","2","Rural","19321.0"
6 "2000","1100031","Cabixi","1","Urbana","2685.0"
7 "2000","1100031","Cabixi","2","Rural","4828.0"
8 "2000","1100049","Cacoal","1","Urbana","51393.0"
9 "2000","1100049","Cacoal","2","Rural","22172.0"
10 "2000","1100056","Cerejeiras","1","Urbana","14832.0"
11 "2000","1100056","Cerejeiras","2","Rural","3357.0"
```

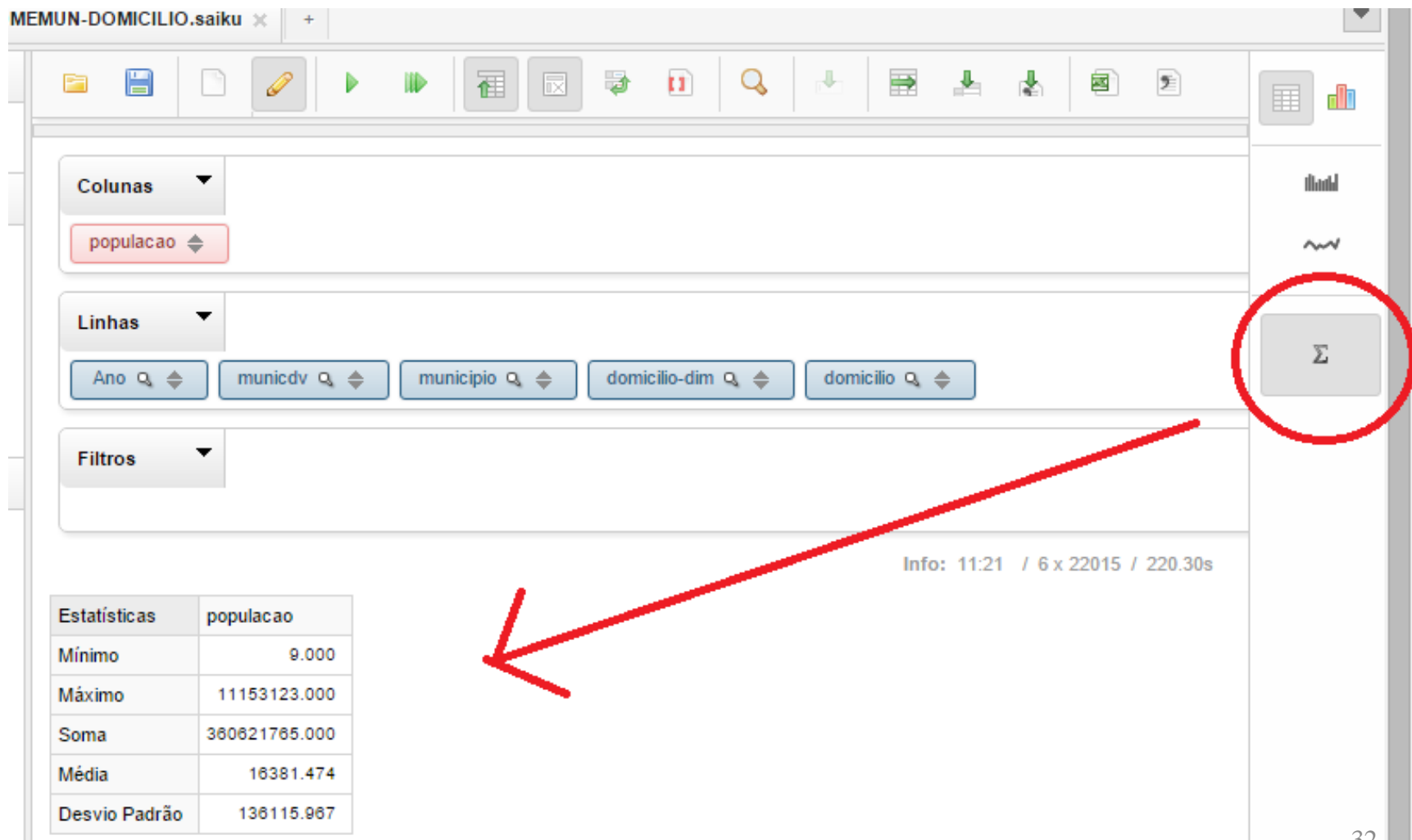
Salvando a consulta

A ferramenta permite salvar a consulta no repositório, para posterior reutilização.



Estatísticas básicas da consulta no Saiku

A ferramenta fornece o recurso de gerar estatísticas básicas sobre os dados da consulta, aplicando funções estatísticas sobre a “medida” observada.



The screenshot displays the Saiku BI tool interface for a query named 'MEMUN-DOMICILIO.saiku'. The interface includes a toolbar with various icons for file operations and data manipulation. The main configuration area is divided into three sections: 'Colunas' (Columns) with 'populacao' selected; 'Linhas' (Rows) with 'Ano', 'municdv', 'municipio', 'domicilio-dim', and 'domicilio' selected; and 'Filtros' (Filters). A red circle highlights the 'Σ' (Sum) button in the right sidebar, and a red arrow points from it to the 'Soma' row in the statistics table. The table shows the following data:

Estatísticas	populacao
Mínimo	9.000
Máximo	11153123.000
Soma	360621765.000
Média	16381.474
Desvio Padrão	136115.967

Info: 11:21 / 6 x 22015 / 220.30s

Atividade

- Construir um cubo para representar o tema “Censo Agropecuário”.
 - Identificar as dimensões e medidas