Projeto Banco de Dados

Pedro Barros Bastos	RA:204481
Gabriel Volpato Giliotti	RA:197569
Rafael Cabral Pili	RA:185999

Filtragem de dados

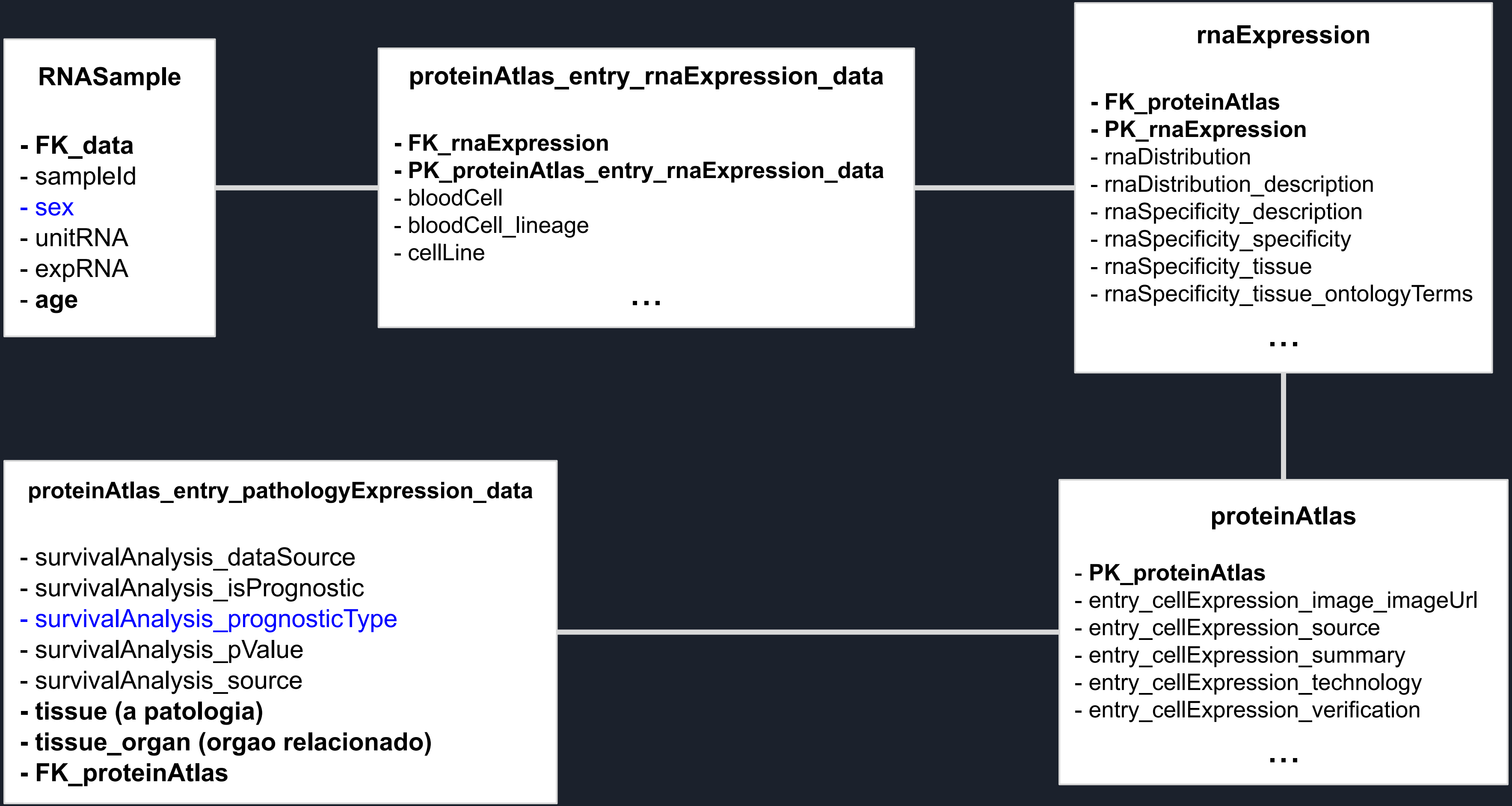
Base: https://www.proteinatlas.org/ENSG00000134057.xml

The Human Protein Atlas is a Swedish-based program initiated in 2003 with the aim to map all the human proteins in cells, tissues and organs using integration of various omics technologies, including antibody-based imaging, mass spectrometry-based proteomics, transcriptomics and systems biology.

- Conversão dos dados da base em XML para formato .TSV para manipulação utilizando o link: https://xmlconverter.sonra.io/signup
- Conversão do .TSV para .CSV para criação de esquemas SQL no Jupyter utilizando o link: https://onlinetsvtools.com/convert-tsv-to-csv
- Melhor entendimento da base dado o modelo Entidade-Relacionamento gerado pela conversão

Problema: (Focado em análise exploratória)

- **Quais patologias possuem amostras de RNA de tecidos afetados dadas por pessoas com mais de X anos?**



RNASample

- **FK_data**
- sampleId
- sex
- unitRNA
- expRNA
- **age**

proteinAtlas_entry_rnaExpression_data

- **FK_rnaExpression**
- **PK_proteinAtlas_entry_rnaExpression_data**
- bloodCell
- bloodCell_lineage
- cellLine

...

rnaExpression

- **FK_proteinAtlas**
- **PK_rnaExpression**
- rnaDistribution
- rnaDistribution_description
- rnaSpecificity_description
- rnaSpecificity_specificity
- rnaSpecificity_tissue
- rnaSpecificity_tissue_ontologyTerms

...

proteinAtlas_entry_pathologyExpression_data

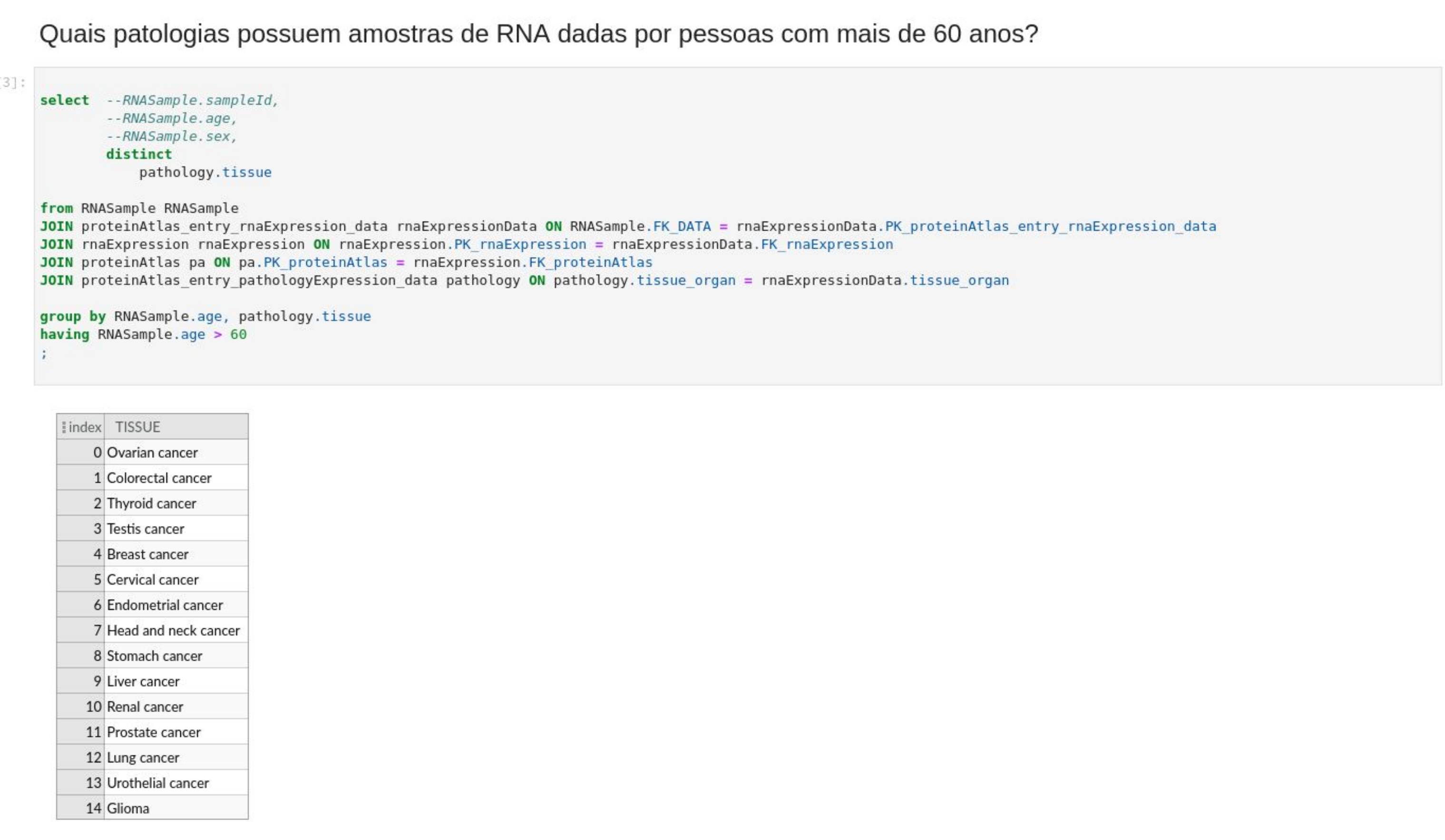
- survivalAnalysis_dataSource
- survivalAnalysis_isPrognostic
- survivalAnalysis_prognosticType
- survivalAnalysis_pValue
- survivalAnalysis_source
- **tissue (a patologia)**
- **tissue_organ (orgao relacionado)**
- **FK_proteinAtlas**

proteinAtlas

- **PK_proteinAtlas**
- entry_cellExpression_image_imageUrl
- entry_cellExpression_source
- entry_cellExpression_summary
- entry_cellExpression_technology
- entry_cellExpression_verification

...

SQL (melhor definição do problema)

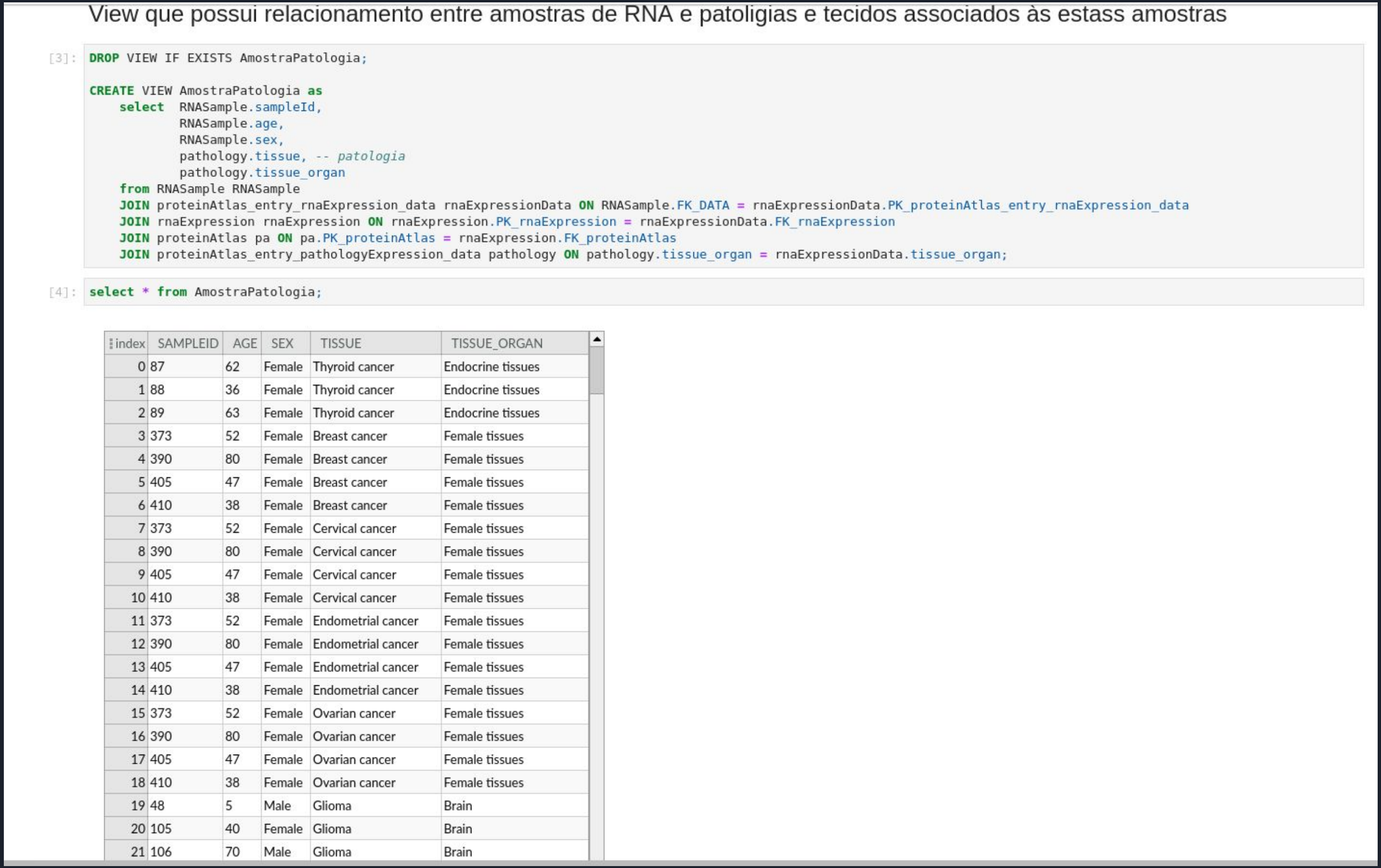


Quais patologias possuem amostras de RNA dadas por pessoas com mais de 60 anos?

3]:

```sql
select  --RNASample.sampleId,
        --RNASample.age,
        --RNASample.sex,
        distinct
            pathology.tissue

from RNASample RNASample
JOIN proteinAtlas_entry_rnaExpression_data rnaExpressionData ON RNASample.FK_DATA = rnaExpressionData.PK_proteinAtlas_entry_rnaExpression_data
JOIN rnaExpression rnaExpression ON rnaExpression.PK_rnaExpression = rnaExpressionData.FK_rnaExpression
JOIN proteinAtlas pa ON pa.PK_proteinAtlas = rnaExpression.FK_proteinAtlas
JOIN proteinAtlas_entry_pathologyExpression_data pathology ON pathology.tissue_organ = rnaExpressionData.tissue_organ

group by RNASample.age, pathology.tissue
having RNASample.age > 60
;
```

index	TISSUE
0	Ovarian cancer
1	Colorectal cancer
2	Thyroid cancer
3	Testis cancer
4	Breast cancer
5	Cervical cancer
6	Endometrial cancer
7	Head and neck cancer
8	Stomach cancer
9	Liver cancer
10	Renal cancer
11	Prostate cancer
12	Lung cancer
13	Urothelial cancer
14	Glioma



View que possui relacionamento entre amostras de RNA e patoligias e tecidos associados às estass amostras

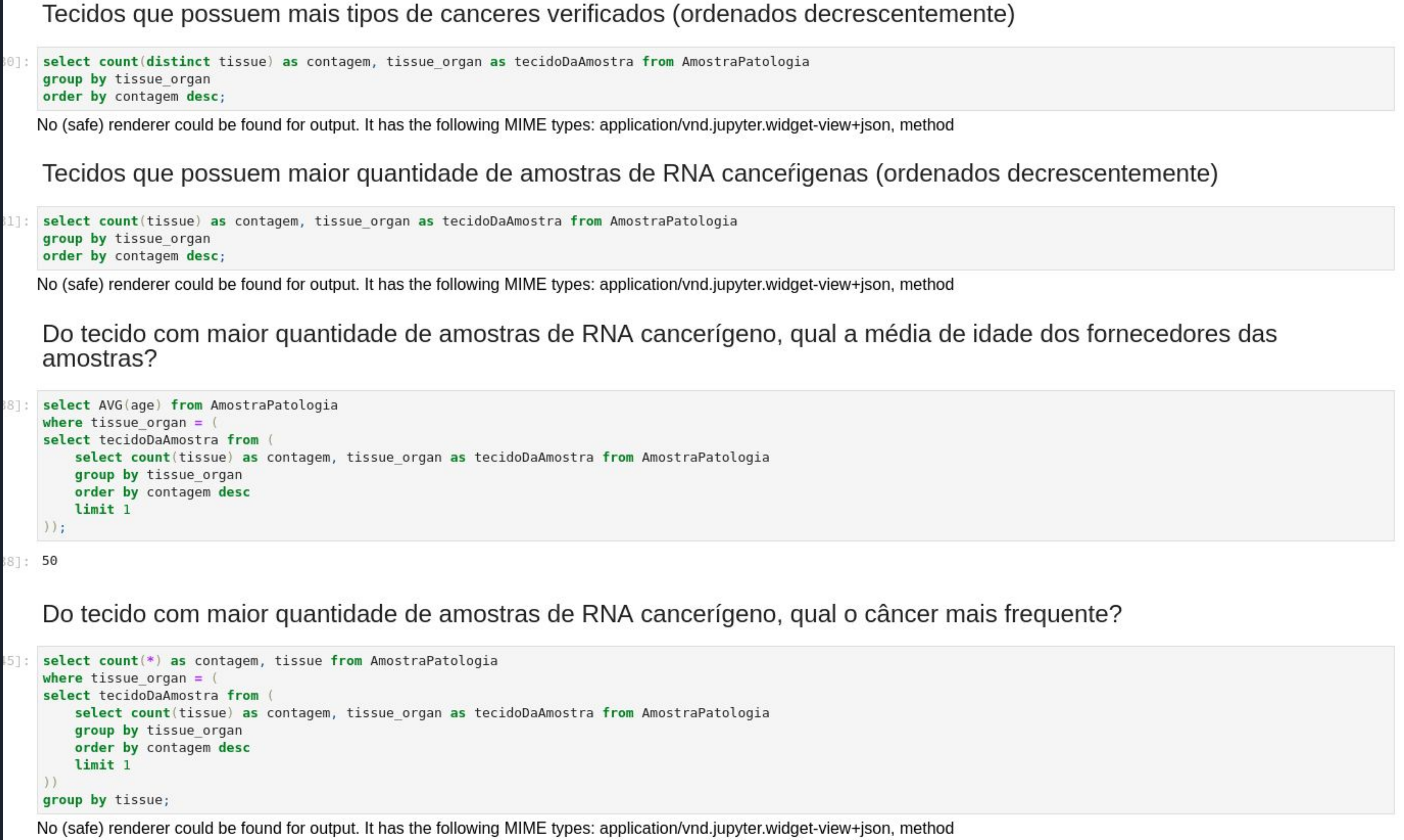
[3]:

```
DROP VIEW IF EXISTS AmostraPatologia;

CREATE VIEW AmostraPatologia as
    select  RNASample.sampleId,
            RNASample.age,
            RNASample.sex,
            pathology.tissue, -- patologia
            pathology.tissue_organ
    from RNASample RNASample
    JOIN proteinAtlas_entry_rnaExpression_data rnaExpressionData ON RNASample.FK_DATA = rnaExpressionData.PK_proteinAtlas_entry_rnaExpression_data
    JOIN rnaExpression rnaExpression ON rnaExpression.PK_rnaExpression = rnaExpressionData.FK_rnaExpression
    JOIN proteinAtlas pa ON pa.PK_proteinAtlas = rnaExpression.FK_proteinAtlas
    JOIN proteinAtlas_entry_pathologyExpression_data pathology ON pathology.tissue_organ = rnaExpressionData.tissue_organ;
```

[4]:

```
select * from AmostraPatologia;
```

index	SAMPLEID	AGE	SEX	TISSUE	TISSUE_ORGAN
0	87	62	Female	Thyroid cancer	Endocrine tissues
1	88	36	Female	Thyroid cancer	Endocrine tissues
2	89	63	Female	Thyroid cancer	Endocrine tissues
3	373	52	Female	Breast cancer	Female tissues
4	390	80	Female	Breast cancer	Female tissues
5	405	47	Female	Breast cancer	Female tissues
6	410	38	Female	Breast cancer	Female tissues
7	373	52	Female	Cervical cancer	Female tissues
8	390	80	Female	Cervical cancer	Female tissues
9	405	47	Female	Cervical cancer	Female tissues
10	410	38	Female	Cervical cancer	Female tissues
11	373	52	Female	Endometrial cancer	Female tissues
12	390	80	Female	Endometrial cancer	Female tissues
13	405	47	Female	Endometrial cancer	Female tissues
14	410	38	Female	Endometrial cancer	Female tissues
15	373	52	Female	Ovarian cancer	Female tissues
16	390	80	Female	Ovarian cancer	Female tissues
17	405	47	Female	Ovarian cancer	Female tissues
18	410	38	Female	Ovarian cancer	Female tissues
19	48	5	Male	Glioma	Brain
20	105	40	Female	Glioma	Brain
21	106	70	Male	Glioma	Brain



Tecidos que possuem mais tipos de canceres verificados (ordenados decrescentemente)

```
select count(distinct tissue) as contagem, tissue_organ as tecidoDaAmostra from AmostraPatologia
group by tissue_organ
order by contagem desc;
```

No (safe) renderer could be found for output. It has the following MIME types: application/vnd.jupyter.widget-view+json, method

Tecidos que possuem maior quantidade de amostras de RNA canceŕigenas (ordenados decrescentemente)

```
select count(tissue) as contagem, tissue_organ as tecidoDaAmostra from AmostraPatologia
group by tissue_organ
order by contagem desc;
```

No (safe) renderer could be found for output. It has the following MIME types: application/vnd.jupyter.widget-view+json, method

Do tecido com maior quantidade de amostras de RNA cancerígeno, qual a média de idade dos fornecedores das amostras?

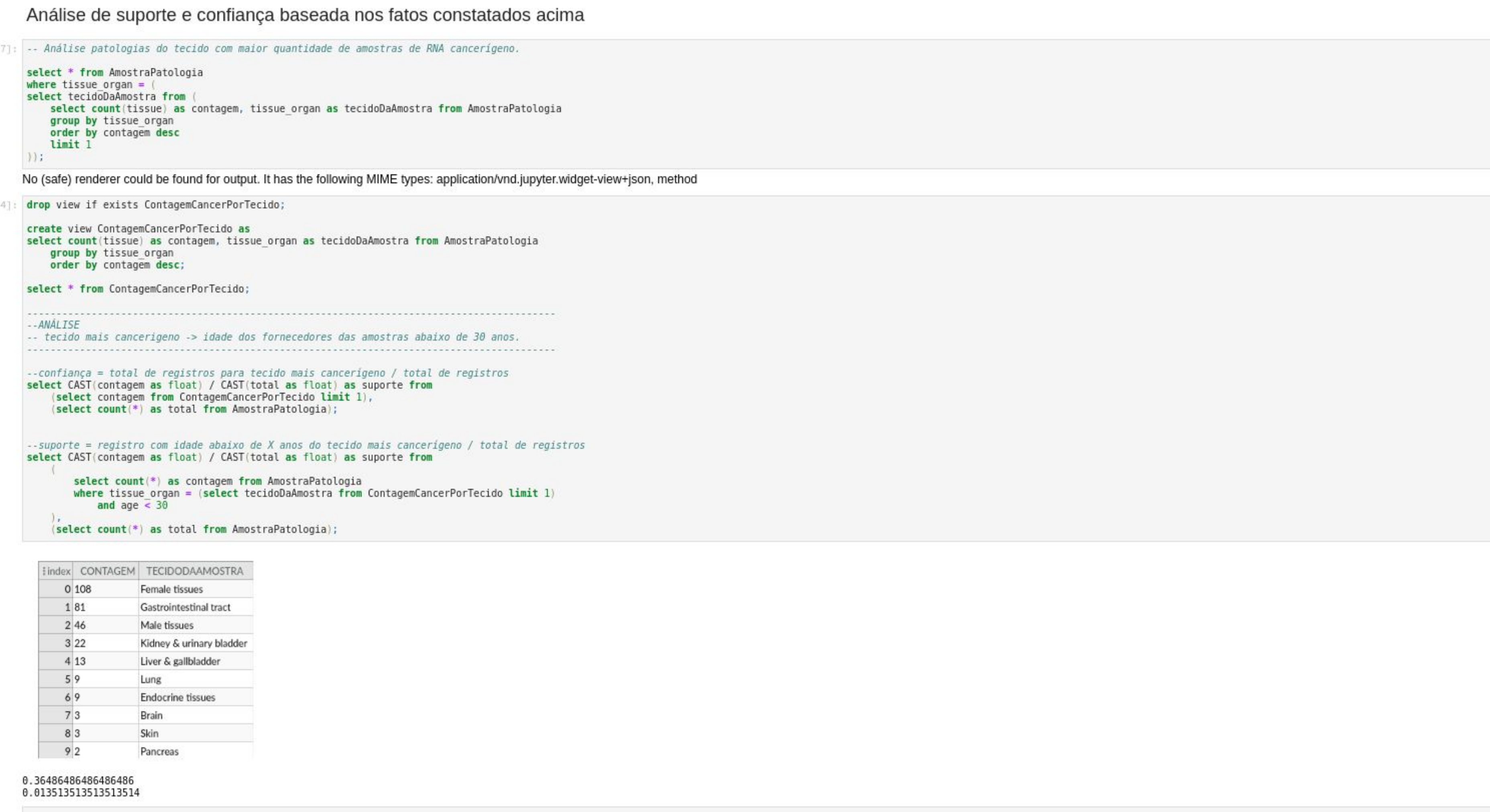
```
select AVG(age) from AmostraPatologia
where tissue_organ = (
select tecidoDaAmostra from (
    select count(tissue) as contagem, tissue_organ as tecidoDaAmostra from AmostraPatologia
    group by tissue_organ
    order by contagem desc
    limit 1
));
```

```
8]: 50
```

Do tecido com maior quantidade de amostras de RNA cancerígeno, qual o câncer mais frequente?

```
select count(*) as contagem, tissue from AmostraPatologia
where tissue_organ = (
select tecidoDaAmostra from (
    select count(tissue) as contagem, tissue_organ as tecidoDaAmostra from AmostraPatologia
    group by tissue_organ
    order by contagem desc
    limit 1
))
group by tissue;
```

No (safe) renderer could be found for output. It has the following MIME types: application/vnd.jupyter.widget-view+json, method



Análise de suporte e confiança baseada nos fatos constatados acima

```
-- Análise patologias do tecido com maior quantidade de amostras de RNA cancerígeno.

select * from AmostraPatologia
where tissue_organ = (
select tecidoDaAmostra from (
    select count(tissue) as contagem, tissue_organ as tecidoDaAmostra from AmostraPatologia
    group by tissue_organ
    order by contagem desc
    limit 1
));
```

No (safe) renderer could be found for output. It has the following MIME types: application/vnd.jupyter.widget-view+json, method

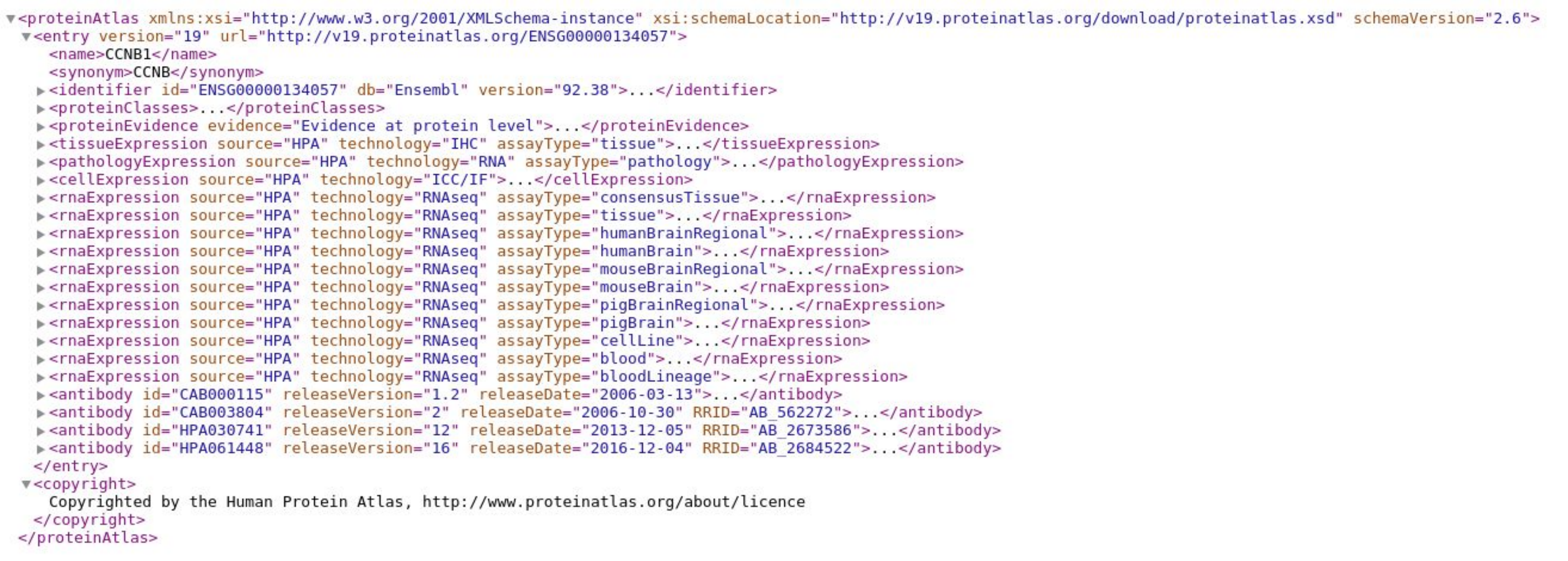
```
drop view if exists ContagemCancerPorTecido;

create view ContagemCancerPorTecido as
select count(tissue) as contagem, tissue_organ as tecidoDaAmostra from AmostraPatologia
    group by tissue_organ
    order by contagem desc;

select * from ContagemCancerPorTecido;

-----------------------------------------------------------------------------------------
--ANÁLISE
-- tecido mais cancerigeno -> idade dos fornecedores das amostras abaixo de 30 anos.
-----------------------------------------------------------------------------------------

--confiança = total de registros para tecido mais cancerígeno / total de registros
select CAST(contagem as float) / CAST(total as float) as suporte from
    (select contagem from ContagemCancerPorTecido limit 1),
    (select count(*) as total from AmostraPatologia);


--suporte = registro com idade abaixo de X anos do tecido mais cancerígeno / total de registros
select CAST(contagem as float) / CAST(total as float) as suporte from
    (
        select count(*) as contagem from AmostraPatologia
        where tissue_organ = (select tecidoDaAmostra from ContagemCancerPorTecido limit 1)
            and age < 30
    ),
    (select count(*) as total from AmostraPatologia);
```

index	CONTAGEM	TECIDODAAMOSTRA
0	108	Female tissues
1	81	Gastrointestinal tract
2	46	Male tissues
3	22	Kidney & urinary bladder
4	13	Liver & gallbladder
5	9	Lung
6	9	Endocrine tissues
7	3	Brain
8	3	Skin
9	2	Pancreas

```
0.36486486486486486
0.013513513513513514
```

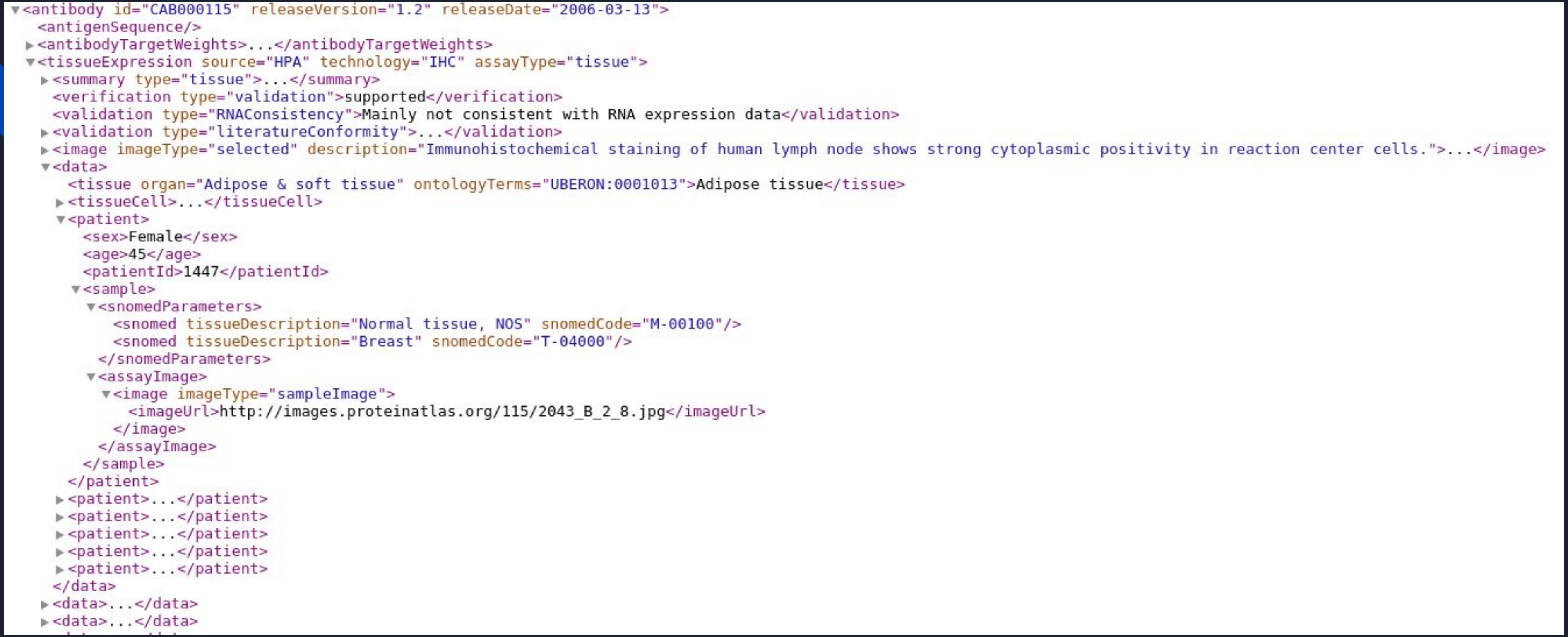
Human Protein Atlas

Base: https://www.proteinatlas.org/ENSG00000134057.xml



```
▼<proteinAtlas xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://v19.proteinatlas.org/download/proteinatlas.xsd" schemaVersion="2.6">
  ▼<entry version="19" url="http://v19.proteinatlas.org/ENSG00000134057">
     <name>CCNB1</name>
     <synonym>CCNB</synonym>
    ▶<identifier id="ENSG00000134057" db="Ensembl" version="92.38">...</identifier>
    ▶<proteinClasses>...</proteinClasses>
    ▶<proteinEvidence evidence="Evidence at protein level">...</proteinEvidence>
    ▶<tissueExpression source="HPA" technology="IHC" assayType="tissue">...</tissueExpression>
    ▶<pathologyExpression source="HPA" technology="RNA" assayType="pathology">...</pathologyExpression>
    ▶<cellExpression source="HPA" technology="ICC/IF">...</cellExpression>
    ▶<rnaExpression source="HPA" technology="RNAseq" assayType="consensusTissue">...</rnaExpression>
    ▶<rnaExpression source="HPA" technology="RNAseq" assayType="tissue">...</rnaExpression>
    ▶<rnaExpression source="HPA" technology="RNAseq" assayType="humanBrainRegional">...</rnaExpression>
    ▶<rnaExpression source="HPA" technology="RNAseq" assayType="humanBrain">...</rnaExpression>
    ▶<rnaExpression source="HPA" technology="RNAseq" assayType="mouseBrainRegional">...</rnaExpression>
    ▶<rnaExpression source="HPA" technology="RNAseq" assayType="mouseBrain">...</rnaExpression>
    ▶<rnaExpression source="HPA" technology="RNAseq" assayType="pigBrainRegional">...</rnaExpression>
    ▶<rnaExpression source="HPA" technology="RNAseq" assayType="pigBrain">...</rnaExpression>
    ▶<rnaExpression source="HPA" technology="RNAseq" assayType="cellLine">...</rnaExpression>
    ▶<rnaExpression source="HPA" technology="RNAseq" assayType="blood">...</rnaExpression>
    ▶<rnaExpression source="HPA" technology="RNAseq" assayType="bloodLineage">...</rnaExpression>
    ▶<antibody id="CAB000115" releaseVersion="1.2" releaseDate="2006-03-13">...</antibody>
    ▶<antibody id="CAB003804" releaseVersion="2" releaseDate="2006-10-30" RRID="AB_562272">...</antibody>
    ▶<antibody id="HPA030741" releaseVersion="12" releaseDate="2013-12-05" RRID="AB_2673586">...</antibody>
    ▶<antibody id="HPA061448" releaseVersion="16" releaseDate="2016-12-04" RRID="AB_2684522">...</antibody>
   </entry>
  ▼<copyright>
     Copyrighted by the Human Protein Atlas, http://www.proteinatlas.org/about/licence
   </copyright>
 </proteinAtlas>
```

Base a partir de sua amostragem hierárquica, partindo do documento em XML



```
▼<antibody id="CAB000115" releaseVersion="1.2" releaseDate="2006-03-13">
   <antigenSequence/>
  ▶<antibodyTargetWeights>...</antibodyTargetWeights>
  ▼<tissueExpression source="HPA" technology="IHC" assayType="tissue">
    ▶<summary type="tissue">...</summary>
     <verification type="validation">supported</verification>
     <validation type="RNAConsistency">Mainly not consistent with RNA expression data</validation>
    ▶<validation type="literatureConformity">...</validation>
    ▶<image imageType="selected" description="Immunohistochemical staining of human lymph node shows strong cytoplasmic positivity in reaction center cells.">...</image>
    ▼<data>
       <tissue organ="Adipose & soft tissue" ontologyTerms="UBERON:0001013">Adipose tissue</tissue>
      ▶<tissueCell>...</tissueCell>
      ▼<patient>
         <sex>Female</sex>
         <age>45</age>
         <patientId>1447</patientId>
        ▼<sample>
          ▼<snomedParameters>
             <snomed tissueDescription="Normal tissue, NOS" snomedCode="M-00100"/>
             <snomed tissueDescription="Breast" snomedCode="T-04000"/>
           </snomedParameters>
          ▼<assayImage>
            ▼<image imageType="sampleImage">
               <imageUrl>http://images.proteinatlas.org/115/2043_B_2_8.jpg</imageUrl>
             </image>
           </assayImage>
         </sample>
       </patient>
      ▶<patient>...</patient>
      ▶<patient>...</patient>
      ▶<patient>...</patient>
      ▶<patient>...</patient>
      ▶<patient>...</patient>
     </data>
    ▶<data>...</data>
    ▶<data>...</data>
```

- Entendimento da hierarquia da base: Descobrir proteínas que podem causar câncer, através da reação com anticorpos
- Problemas propostos:
    - Percentual de células cancerígenas dentre todos os anticorpos (total de tumores/total de amostras)
    - Percentual de células cancerígenas para cada anticorpo (total de tumores para cada anticorpo/total de amostras do anticorpo com tumores)

Análises e Resultados obtidos:

```
let $protein := doc('http://www.proteinatlas.org/ENSG00000134057.xml')

let $totalTissue := ($protein//proteinAtlas/entry/antibody/tissueExpression/data/tissueCell)
let $totalDeTecidos := count($totalTissue)

let $totalTumorTissue:=
($protein//proteinAtlas/entry/antibody/tissueExpression/data/tissueCell[contains(cellType/text(),'Tumor')])
let $totalDeTumores := count($totalTumorTissue)

for $c in ($protein//proteinAtlas/entry/antibody)
let $qtdTumoresPorAnticorpo := $c//tissueExpression/data/tissueCell[contains(cellType/text(),'Tumor')]
return count($qtdTumoresPorAnticorpo) div ($totalDeTecidos)*100
```

```
6.75675%
```

```
for $c in ($protein//proteinAtlas/entry/antibody)
let $qtdTumoresPorAnticorpo := $c//tissueExpression/data/tissueCell[contains(cellType/text(),'Tumor')]
return count($qtdTumoresPorAnticorpo) div ($totalDeTumores)*100
```

```
CAB000115   -->  33.3 %
CAB003804   -->  33.3%
HPA030741   -->  0.0%
HPA061448   -->  33.3%
```

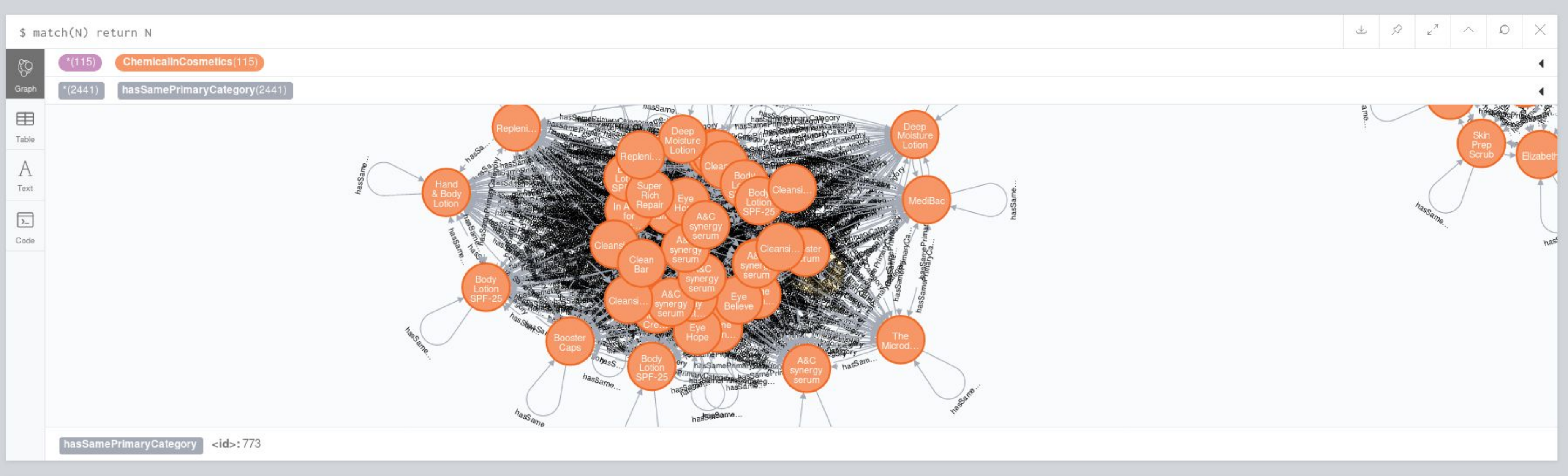
```
let $totalRna:= count($protein//proteinAtlas/*/rnaExpression/data)

let $x := for $c in ($protein//proteinAtlas/entry/rnaExpression)
where $c/data/*[@expRNA > 60][@expRNA < 200][@type="RNAExpression"]
return $c
let $tot := for $c in ($protein//proteinAtlas/entry/rnaExpression/data)
where $c/tissue[@organ='Endocrine tissues']
return $c

let $y := for $c in ($protein//proteinAtlas/entry/rnaExpression/data)
where $c/*[@expRNA < 15][@type="RNAExpression"] and
$c/tissue[@organ='Endocrine tissues']
return $c
return count($y) div count($tot)

let $z := for $c in ($protein//proteinAtlas/entry/rnaExpression/data)
where $c/*[@expRNA > 10][@type="RNAExpression"] and
count($c/*[@sex="Female"]) > count($c/*[@sex="Male"])
return $c
```

Utilizando XQuery foi muito mais simples realizar a busca por dados hierarquizados que em SQL. Especialmente, como a hierarquia possui muitos níveis a XQuery ajudou muito a recuperar os dados de interesse. Por exemplo, como a hierarquia é grande demais para compreender todos seus dados, através da XQuery ficou mais simples filtrar os campos de interesse onde era necessário buscar atributos específicos em diferentes e indefinidas nós. Em SQL, onde seria necessário realizar indefinidos números de JOINs.

Alguns resultados interessantes:

- Foram encontrados 162 tecidos cancerígenos com RNAm expressivo.
- Alguns tecidos como o Endócrino e o Cerebral apresentam altas quantidades de RNAm quando estão com câncer.
- Tecidos da tireóide, colo, e pele cancerígenos apresentam maior recorrência em mulheres do que em homens, enquanto homens apresentam mais casos de câncer de reto e intestino que mulheres.

Cypher



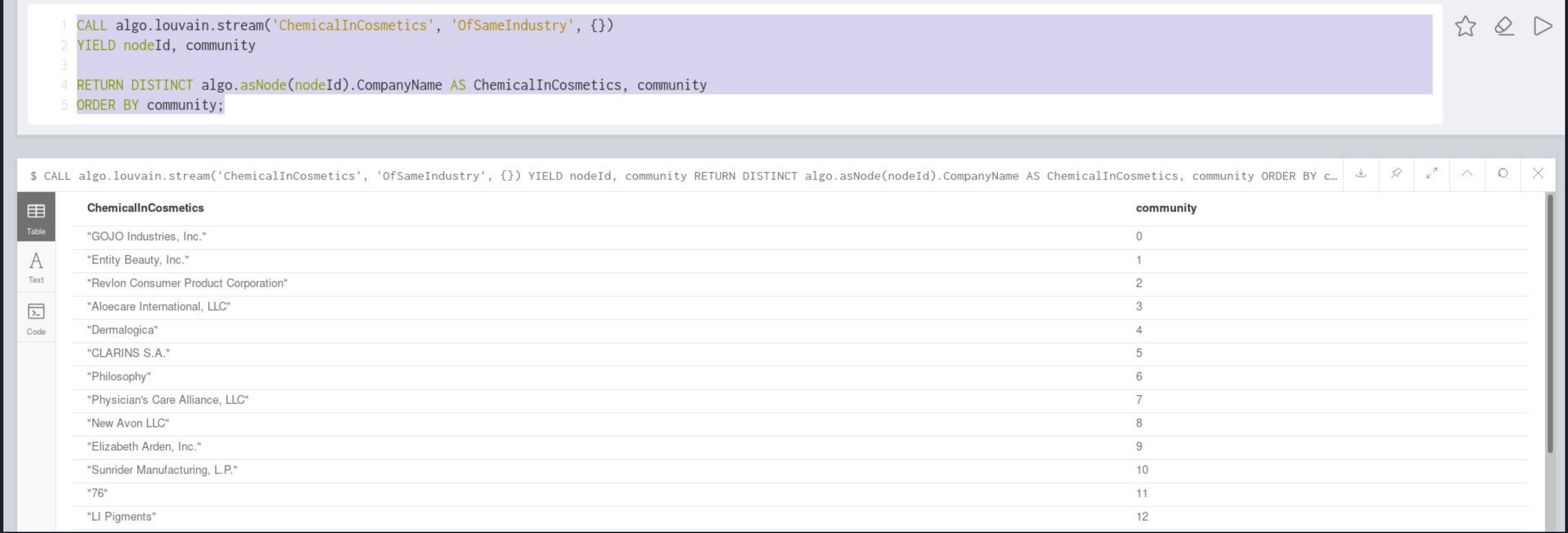
```
$ match(N) return N
```

*(115) ChemicalInCosmetics(115)

*(2441) hasSamePrimaryCategory(2441)

hasSamePrimaryCategory <id>: 773

Análise de produtos cancerígenos verificando comunidades baseadas por indústria.

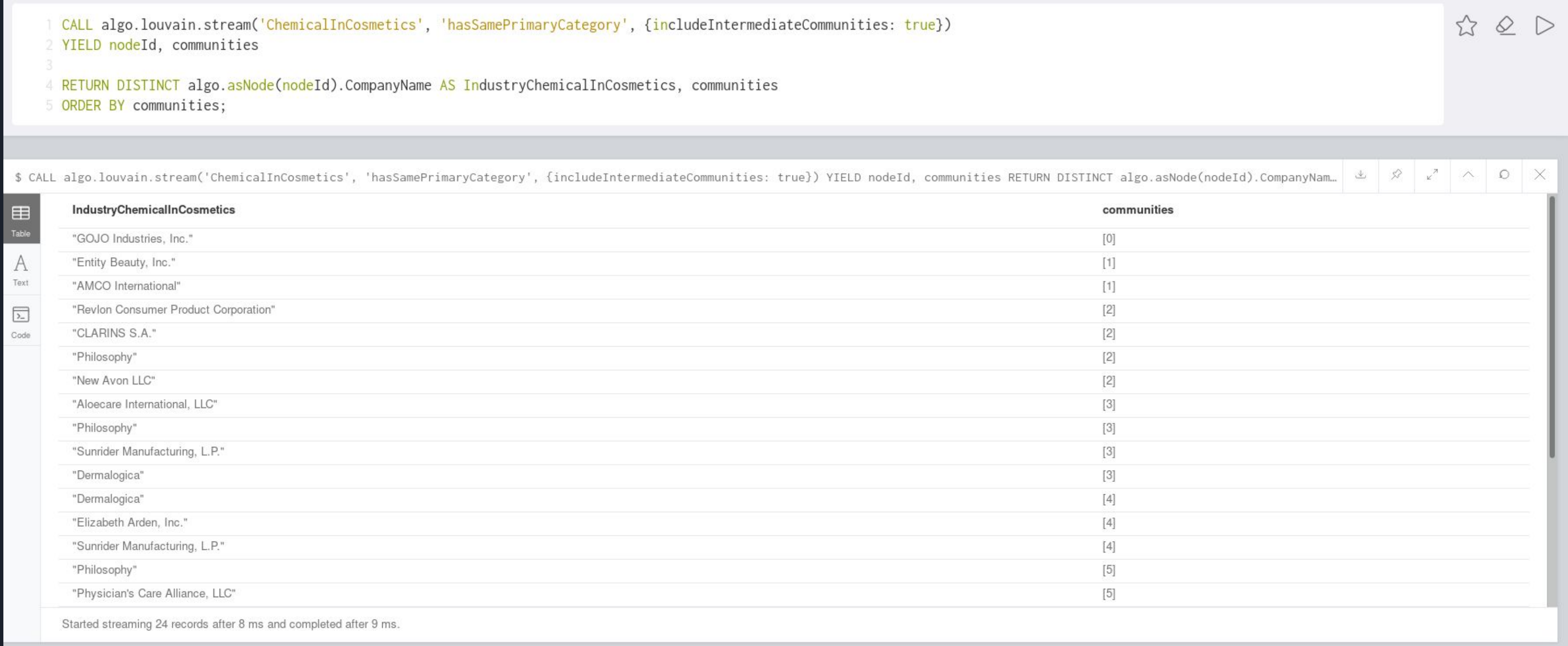


```
CALL algo.louvain.stream('ChemicalInCosmetics', 'OfSameIndustry', {})
YIELD nodeId, community

RETURN DISTINCT algo.asNode(nodeId).CompanyName AS ChemicalInCosmetics, community
ORDER BY community;
```

```
$ CALL algo.louvain.stream('ChemicalInCosmetics', 'OfSameIndustry', {}) YIELD nodeId, community RETURN DISTINCT algo.asNode(nodeId).CompanyName AS ChemicalInCosmetics, community ORDER BY c…
```

ChemicalInCosmetics	community
"GOJO Industries, Inc."	0
"Entity Beauty, Inc."	1
"Revlon Consumer Product Corporation"	2
"Aloecare International, LLC"	3
"Dermalogica"	4
"CLARINS S.A."	5
"Philosophy"	6
"Physician's Care Alliance, LLC"	7
"New Avon LLC"	8
"Elizabeth Arden, Inc."	9
"Sunrider Manufacturing, L.P."	10
"76"	11
"LI Pigments"	12

Análise de produtos cancerígenos verificando comunidades associadas por categoria primária do produto.



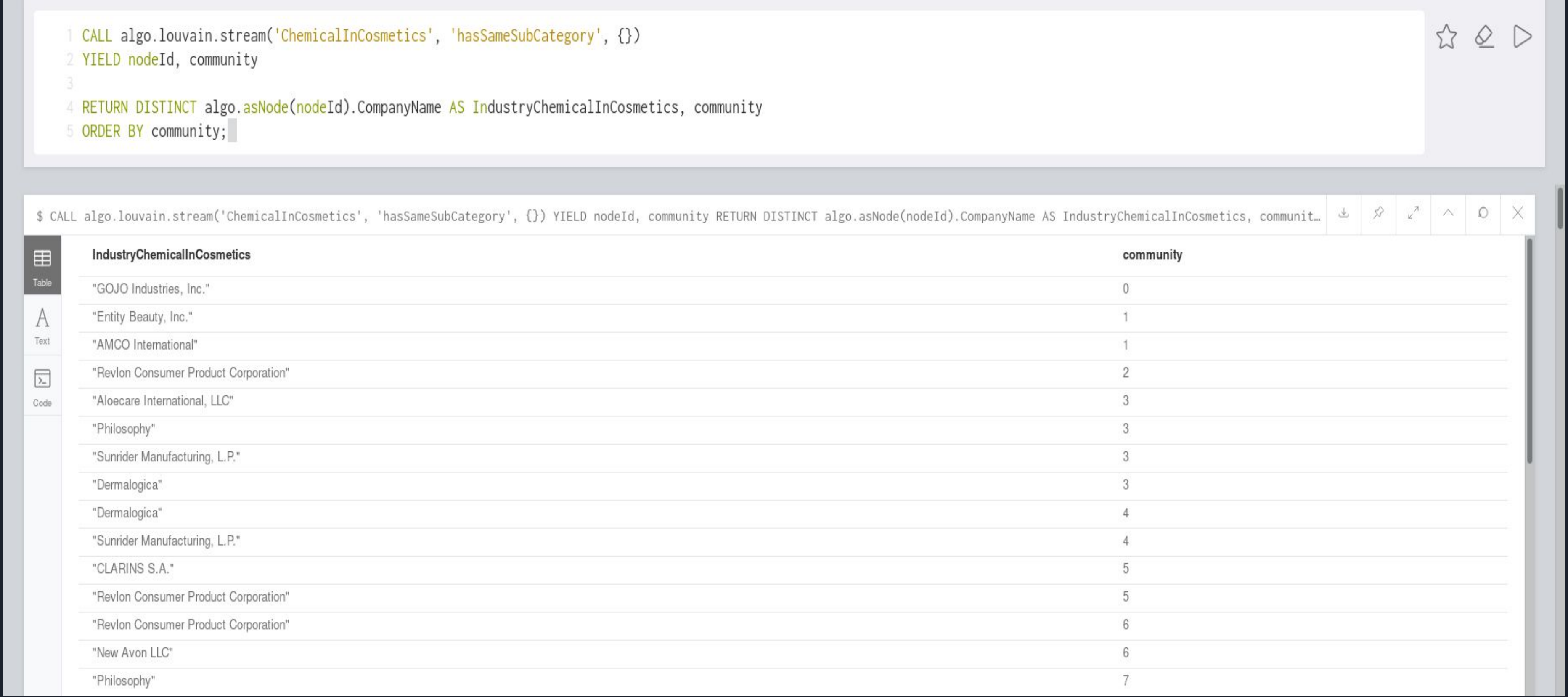
```
CALL algo.louvain.stream('ChemicalInCosmetics', 'hasSamePrimaryCategory', {includeIntermediateCommunities: true})
YIELD nodeId, communities

RETURN DISTINCT algo.asNode(nodeId).CompanyName AS IndustryChemicalInCosmetics, communities
ORDER BY communities;
```

```
$ CALL algo.louvain.stream('ChemicalInCosmetics', 'hasSamePrimaryCategory', {includeIntermediateCommunities: true}) YIELD nodeId, communities RETURN DISTINCT algo.asNode(nodeId).CompanyNam…
```

IndustryChemicalInCosmetics	communities
"GOJO Industries, Inc."	[0]
"Entity Beauty, Inc."	[1]
"AMCO International"	[1]
"Revlon Consumer Product Corporation"	[2]
"CLARINS S.A."	[2]
"Philosophy"	[2]
"New Avon LLC"	[2]
"Aloecare International, LLC"	[3]
"Philosophy"	[3]
"Sunrider Manufacturing, L.P."	[3]
"Dermalogica"	[3]
"Dermalogica"	[4]
"Elizabeth Arden, Inc."	[4]
"Sunrider Manufacturing, L.P."	[4]
"Philosophy"	[5]
"Physician's Care Alliance, LLC"	[5]

Started streaming 24 records after 8 ms and completed after 9 ms.

Análise de produtos cancerígenos verificando comunidades associadas por subcategorias do produto.



```
CALL algo.louvain.stream('ChemicalInCosmetics', 'hasSameSubCategory', {})
YIELD nodeId, community

RETURN DISTINCT algo.asNode(nodeId).CompanyName AS IndustryChemicalInCosmetics, community
ORDER BY community;
```

```
$ CALL algo.louvain.stream('ChemicalInCosmetics', 'hasSameSubCategory', {}) YIELD nodeId, community RETURN DISTINCT algo.asNode(nodeId).CompanyName AS IndustryChemicalInCosmetics, communit…
```

IndustryChemicalInCosmetics	community
"GOJO Industries, Inc."	0
"Entity Beauty, Inc."	1
"AMCO International"	1
"Revlon Consumer Product Corporation"	2
"Aloecare International, LLC"	3
"Philosophy"	3
"Sunrider Manufacturing, L.P."	3
"Dermalogica"	3
"Dermalogica"	4
"Sunrider Manufacturing, L.P."	4
"CLARINS S.A."	5
"Revlon Consumer Product Corporation"	5
"Revlon Consumer Product Corporation"	6
"New Avon LLC"	6
"Philosophy"	7