

UNIVERSIDADE FEDERAL DE MINAS GERAIS
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

Introdução a Banco de Dados
Trabalho Prático II – À Procura da Felicidade

Artur Marzano
João Marcos Couto
João Victor Tamm
Tiago de Rezende Alves

Professora: Mirella M. Moro
Monitor:

Belo Horizonte
26 de novembro de 2018

Sumário

1	Introdução	2
2	Dados utilizados	2
2.1	Filtragem dos dados	4
3	Diagrama ER	4
4	Diagrama Relacional	4
5	Consultas realizadas	5
6	Consultas do tipo relatório	8
7	Consultas com dados não convencionais	11
8	Interface	11
9	Atividades do grupo	13
10	Conclusão	14
11	Referências Bibliográficas	14

1 Introdução

Neste trabalho, foi modelado um banco de dados relacional utilizando dois *datasets* disponibilizados pelo portal de ciência de dados Kaggle (www.kaggle.com), com o intuito de possibilitar a realização de consultas interessantes variadas. As consultas foram criadas de modo a encontrar relações entre o desenvolvimento econômico de um país e o índice de felicidade de seus habitantes.

2 Dados utilizados

Os dados utilizados no trabalho advêm de dois *datasets* diferentes. O primeiro deles, "World Development Indicators", foi criado pela Instituição financeira World Bank e tem como objetivo principal armazenar 1345 indicadores de desenvolvimento de um total de 247 países ao longo de 56 anos (1960 a 2015). Para tal, ele é composto por um total de 6 tabelas. A primeira delas, Country, retém uma série de dados gerais sobre cada um dos países indexados, como por exemplo o ano do último censo populacional realizado e a qual região geográfica cada país pertence.

A segunda tabela, CountryNotes, estabelece uma relação entre os países e cada um dos indicadores já que contém informações sobre a origem dos dados em cada um dos países.

A terceira tabela, Footnotes, faz algo similar à anterior porém aqui, as informações de origem dos dados são vinculadas não apenas a países mas também aos anos específicos.

A quarta tabela, Indicators, é a mais extensa pois armazena o valor numérico de cada um dos indicadores de desenvolvimento de cada país nos anos em que estes dados estiveram disponíveis.

A quinta tabela, Series, contém informações sobre cada um dos 1345 indicadores presentes no *dataset*, tal como a definição semântica de cada um deles e a periodicidade com a qual a coleta de cada um é realizada pelos órgãos internacionais responsáveis.

Por fim, a sexta tabela, SeriesNotes, contém comentários diversos sobre indicadores em anos específicos. Estes comentários mencionam, por exemplo, quando um determinado indicador em um certo ano foi interpolado utilizando os valores dos anos imediatamente anteriores, ou, se o valor presente em certa linha de Indicators foi coletada apenas utilizando dados do setor de manufaturados. É válido ressaltar que aqui, as anotações não são atreladas a um país específico.

Após a descrição das tabelas, é pertinente pontuar o que cada um dos dados presentes nelas representam. Os detalhes são apresentados abaixo:

1 Country

- *CountryCode*: Campo único que identifica o país. Ex.: BRA, USA, etc.
- *ShortName*: Campo indica um nome curto para o país. Países que possuem nomes longos como United States of America tem seu nome encurtado para United States.
- *LongName*: Campo registra o nome completo do país, independente do tamanho que ele tenha.
- *Region*: Registra qual a região do globo o país está localizado. Ex.: North America, Europe & Central Asia, South America, etc.
- *IncomeGroup*: Determina qual o nível de arrecadação de um país. Países são divididos em quatro grupos: low, lower-middle, upper-middle e high.
- *OtherGroups*: Indica outro grupos que o país pertence, no banco de dados resultante pode ser *HIPC* ou *Euro area*.
- *LendingCategory*: Determina qual o nível de arrecadação de um país. Países são divididos em quatro grupos: low, lower-middle, upper-middle e high.
- *IMFDataDisseminationStandard*: Armazena informações referentes a disseminação dos dados. No banco de dados resultante assume o valor de General Data Dissemination System (GDDS) ou Special Data Dissemination Standard (SDDS).
- *LatestHouseholdSurvey*: Campo registra o ano em que foi realizada uma pesquisa sobre condições de moradia, quantidade de eletrodomésticos, quartos, etc.

- *LastestPopulationCensus*: Campo registra o ano em que o censo populacional foi realizado pela última vez.
- *LatestAgriculturalCensus*: Campo registra o ano em que o censo agropecuário foi realizado pela última vez.
- *LatestIndustrialData*: Armazena o ano em que os dados de produção de bens de consumo foram obtidos pela última vez.
- *CurrencyUnit*: Moeda utilizada no país.
- *Flag*: Bandeira do país.
- *GeoLat*: Latitude do país.
- *GeoLong*: Longitude do país.

2 CountryNotes

- *SeriesCode*: Código da série associada a descrição.
- *CountryCode*: Código do país associado a descrição.
- *Description*: Descreve o que a série significa no contexto daquele país e como os dados foram obtidos.

3 Footnotes

- *SeriesCode*: Código da série associada as notas de rodapé.
- *CountryCode*: Código do país associado as notas de rodapé.
- *Description*: Descreve como os dados foram obtidos.
- *Year*: Armazena o ano em que os dados foram coletados.

4 Indicators

- *CountryCode*: Código do país associado ao indicador.
- *Year*: Ano que o indicador está associado.
- *IndicatorName*: Campo indica o nome do indicador em questão. Ex.: Adolescent fertility rate (births per 1,000 women ages 15-19).
- *CountryName*: Nome do país associado ao indicador.
- *IndicatorCode*: Campo que determina qual é o indicador.
- *Value*: Campo que armazena o valor do indicador.

5 Series

- *SeriesCode*: Campo único que identifica a série. Ex.: BN.KLT.DINV.CD, etc.
- *Topic*: Campo indica qual o tópico abordado pela série.
- *LongDefinition*: Fornece uma definição completa da série em questão.
- *AggregationMethod*: Registra qual o método de agregação foi utilizado para realizar o cálculo da série.
- *Limitations and Exceptions*: Indica as limitações e excessões existentes devido à deficiência nos dados.
- *GeneralComments*: Comentário geral acerca dos dados utilizados como por exemplo o período do estudo.
- *Source*: Campo registra a instituição que forneceu os dados para o estudo.
- *DevelopmentRelevance*: Representa a relevância que a série tem como indicador de desenvolvimento do país.
- *IndicatorName*: Armazena o nome do indicador cuja série está associada.

6 SeriesNotes

- *SeriesCode*: Campo único que identifica a série. Ex.: BN.KLT.DINV.CD, etc.
- *Year*: Indica os anos que a série abrange.
- *Description*: Contém informações sobre a aquisição dos dados.

Aliado ao dataset anterior, utilizou-se também o dataset "*World Happiness Report*" criado pela *Sustainable Development Solutions Network*. Esses dados tem como função ilustrar o nível de felicidade de cada país, estimados por meios de seis fatores principais: produção econômica, assistência social, expectativa de vida, liberdade, ausência de corrupção e generosidade. Os dados estavam organizados originalmente por meio de três tabelas com a mesma estrutura, uma para

cada ano, a saber, 2015, 2016 e 2017, com informações de 158, 157 e 155 países respectivamente, o que corresponde ao número de linhas da tabela. Desse modo, seus campos serão explicados a seguir:

- *Country*: Nome do país.
- *Region*: Região em que o país pertence.
- *Happiness Rank*: Colocação dos países considerando a felicidade.
- *Happiness Score*: Uma medida adquirida ao perguntar para as pessoas: "Como você classifica sua felicidade de 0 a 10 onde 10 é mais feliz."
- *Standard Error*: Desvio padrão da medida.
- *Economy*: A influência que o GDP tem no cálculo da felicidade.
- *Family*: A influência que a família tem no cálculo da felicidade.
- *Health*: A influência que a expectativa de vida tem no cálculo da felicidade.
- *Freedom*: A influência que a liberdade tem no cálculo da felicidade.
- *Trust*: A influência que a confiança no governo tem no cálculo da felicidade.
- *Generosity*: A influência que a generosidade tem no cálculo da felicidade.
- *Dystopia Residual*: A influência que o distopia residual tem no cálculo da felicidade.

2.1 Filtragem dos dados

Com o intuito de promover uma melhor coesão entre os dois *datasets* utilizados, foi feita uma filtragem dos dados e um *dataset* resultante foi criado, contendo apenas os indicadores de 2015, visto que o *dataset* correspondente aos indicadores de desenvolvimento abrange o intervalo de 1960 até 2015 enquanto o de felicidade foram mantidos e vão de 2015 até 2017. Além da filtragem por data, foi feita a filtragem por país, tendo em vista que existem países que não são comuns aos dois *datasets* escolhidos.

3 Diagrama ER

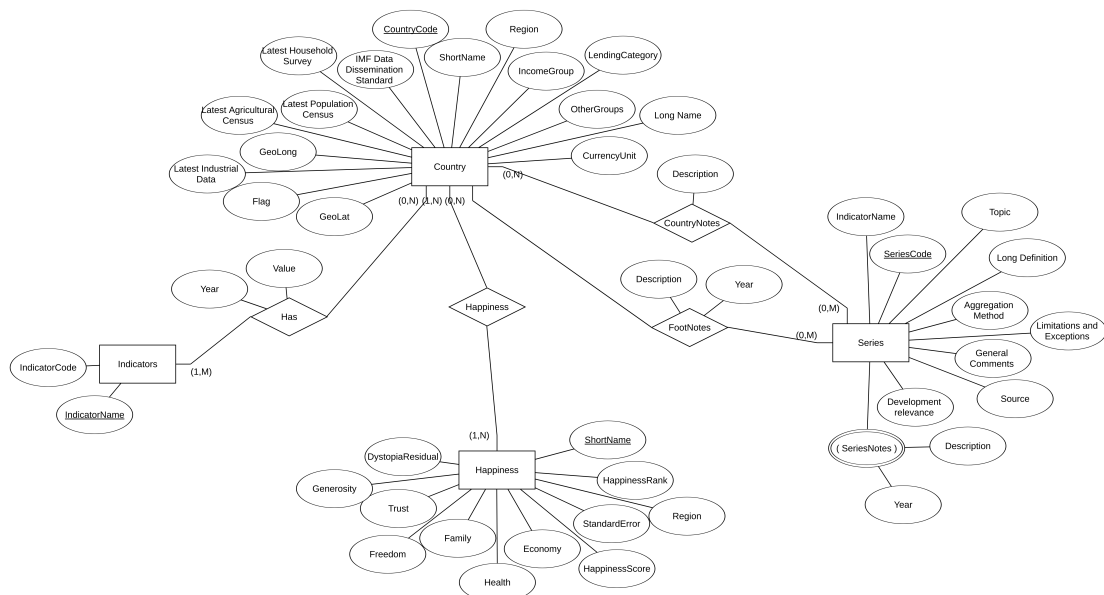


Figura 1 Diagrama ER* do dataset utilizado.

*Os três anos de Happiness foram unificados em uma única entidade por serem iguais

4 Diagrama Relacional

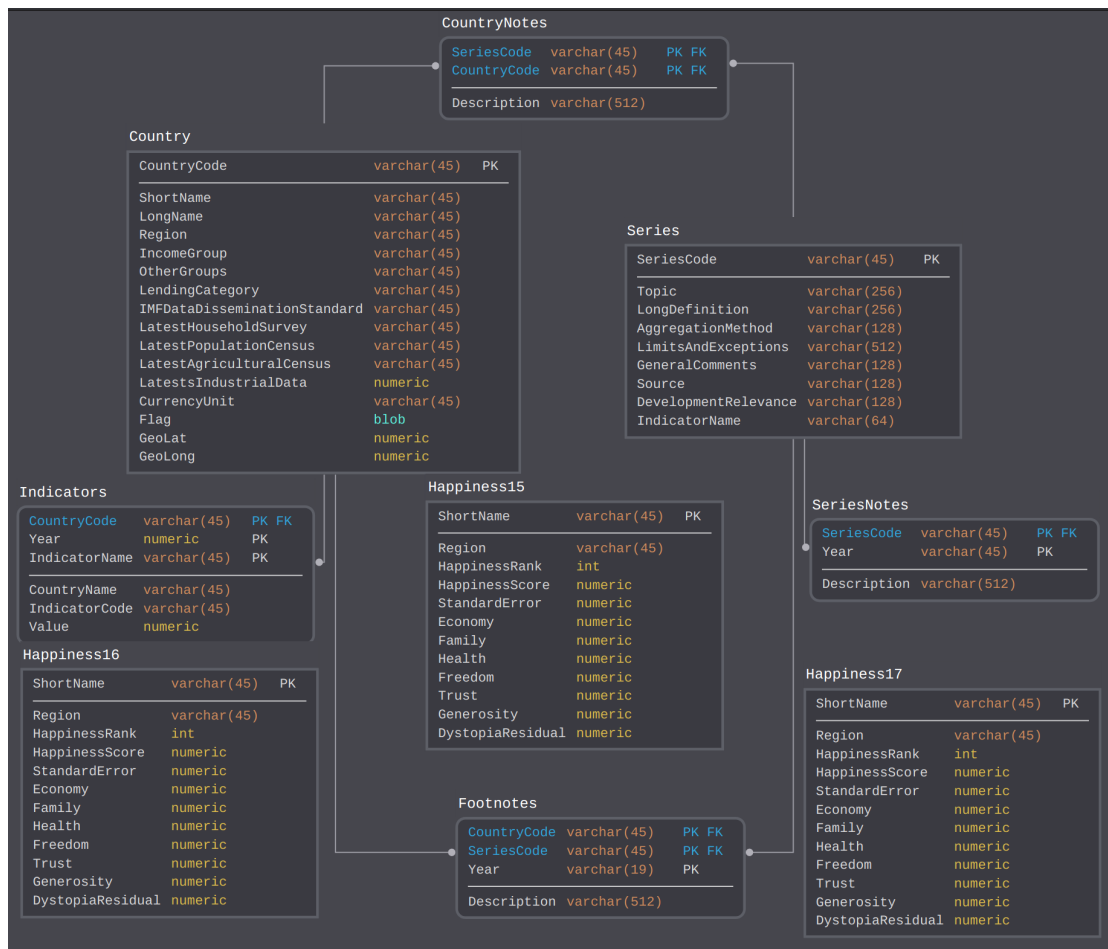


Figura 2 Diagrama relacional do dataset utilizado.

5 Consultas realizadas

- Os 10 países mais felizes dentre aqueles com expectativa de vida menor que a média.

```
>> SELECT CountryName FROM Happiness17 WHERE Health < (SELECT AVG(Health)
FROM Happiness17) ORDER BY HappinessScore DESC LIMIT 10;
```

CountryName
Guatemala
Trinidad and Tobago
Uzbekistan
Russia
Belize
Bolivia
Turkmenistan
Libya
Philippines
Kosovo

- Os 5 países que mais confiam no governo dentre os que pertencem aos 20 mais felizes.

```
>> SELECT H1.CountryName FROM Happiness17 H1 INNER JOIN (SELECT CountryName
FROM Happiness17 ORDER BY HappinessScore DESC LIMIT 20) H2 ON
```

H1.CountryName = H2.CountryName LIMIT 5;

CountryName
Norway
Denmark
Iceland
Switzerland
Finland

3. Felicidade média, mínima e máxima dos países nos quais menos de 1/5 da população tem acesso a serviços sanitários básicos.

```
>> SELECT ROUND(AVG(HappinessScore),4) Avg, MIN(HappinessScore) Min,
      MAX(HappinessScore) Max FROM Happiness15 H NATURAL JOIN Indicators WHERE
      IndicatorCode='SH.STA.ACSN' AND Value < 20;
```

Avg	Min	Max
3.8529	2.839	4.633

4. O nome, índice de felicidade e porcentagem de mulheres no parlamento dos países com felicidade maior que 4 e com as 10 maiores porcentagens de mulheres no parlamento.

```
>> SELECT CountryName, H.HappinessScore, ROUND(I.Value,4) '%women' FROM
      Happiness15 H NATURAL JOIN Indicators I WHERE I.IndicatorCode=
      'SG.GEN.PARL.ZS' AND H.HappinessScore > 4 ORDER BY I.Value DESC LIMIT 10;
```

CountryName	HappinessScore	%women
Bolivia	5.89	53.0769
Sweden	7.364	43.5530
South Africa	4.642	41.9192
Ecuador	5.975	41.6058
Finland	7.406	41.5000
Nicaragua	5.828	41.3043
Iceland	7.561	41.2698
Spain	6.329	41.1429
Norway	7.522	39.6450
Mozambique	4.971	39.6000

5. O nome dos países e o ranking de felicidade, comparado com a confiança no governo, aliada a taxa de impostos e a sua quantidade.

```
>>SELECT H.CountryName, H.HappinessRank, ROUND((Trust / HappinessScore),4)
      Trust, A.Value TaxNum, B.Value TaxRate FROM ((Happiness15 H JOIN
      Indicators A ON H.CountryName = A.CountryName) JOIN Indicators B ON
      A.CountryName = B.CountryName) WHERE A.IndicatorCode='IC.TAX.PAYM' AND
      B.IndicatorCode = 'IC.TAX.TOTL.CP.ZS' ORDER BY Trust DESC;
```

CountryName	HappinessRank	Trust	Taxnum	TaxRate
Rwanda	154	0.1593	25	33
Georgia	130	0.0892	5	16.4
Qatar	28	0.0790	4	11.3
Singapore	24	0.0724	6	18.4
Djibouti	126	0.0643	36	37.6
Denmark	3	0.0642	10	24.5
Sweden	8	0.0595	6	49.1
New Zealand	9	0.0589	8	34.3
Finland	6	0.0559	8	37.9
United Arab Emirates	20	0.0559	4	15.9
...

6. A relevância da economia na felicidade, em conjunto com o acesso ao crédito agrupado por moeda.

```
>> SELECT CurrencyUnit, ROUND(AVG(Economy/HappinessScore),4) Avg, Avg(Value)
Credit FROM ((Country C JOIN Happiness15 H ON C.ShortName = H.CountryName)
JOIN Indicators I ON H.CountryName = I.CountryName) WHERE IndicatorCode =
'IC.CRD.INFO.XQ' GROUP BY CurrencyUnit ORDER BY Avg DESC;
```

CurrencyUnit	Avg	Credit
Qatari riyal	0.2557	5
Kuwaiti dinar	0.2469	6
Bulgarian lev	0.2400	5
Hungarian forint	0.2335	5
Botswana pula	0.2294	6
Singapore dollar	0.2239	7
Bahraini dinar	0.2221	7
Saudi Arabian riyal	0.2177	8
Japanese yen	0.2122	6
Lebanese pound	0.2120	6
...

7. O nome, índice de felicidade, tempo necessário para se iniciar um negócio em dias e índice de impostos dos 10 países mais felizes.

```
>> SELECT H.CountryName, H.HappinessScore, A.Value OpenBusinessTime, B.Value
TaxRate FROM Happiness15 H JOIN Indicators A ON A.CountryName =
H.CountryName JOIN Indicators B ON B.CountryName = A.CountryName WHERE
A.IndicatorCode = 'IC.TAX.DURS' AND B.IndicatorCode='IC.REG.DURS' ORDER BY H.HappinessScore
```

CountryName	HappinessScore	OpenBusinessTime	TaxRate
Switzerland	7.587	63	10
Iceland	7.561	140	4
Denmark	7.527	130	3
Norway	7.522	83	4
Canada	7.427	131	1.5
Finland	7.406	93	14
Netherlands	7.378	123	4
Sweden	7.364	122	7
New Zealand	7.286	152	0.5
Australia	7.284	105	2.5

8. Média por região das médias temporais de felicidade de cada país.

```
>> SELECT C.Region, ROUND(AVG(HIT.AvgScore),4) AvgHappiness FROM
      HappinessInTime HIT JOIN Country C ON C.ShortName = HIT.Country
      GROUP BY C.Region;
```

Region	AvgHappiness
East Asia & Pacific	5.6859
Europe & Central Asia	5.9215
Latin America & Caribbean	6.0509
Middle East & North Africa	5.5874
North America	7.2272
South Asia	4.5908
Sub-Saharan Africa	4.1034

9. A quantidade de países e média da mortalidade infantil por patamar de felicidade:

```
>> SELECT COUNT(CountryName) Count, ROUND(AVG(Value),4) DeathInf,
      ROUND(HappinessScore,0) Patamar FROM Indicators NATURAL JOIN
      Happiness15 WHERE IndicatorCode='SH.DYN.MORT' GROUP BY Patamar;
```

Count	DeathInf	Patamar
4	75.3250	3
26	67.4423	4
45	36.4244	5
33	15.0909	6
26	6.4885	7
4	3.0000	8

10. A quantidade de países e o máximo de felicidade em cada grupo de países, quando agrupados por extratos de imposto total.

```
>> SELECT COUNT(*) Countries, MAX(HappinessScore) MaxHappiness,
      ROUND(Value/10,0)*10 TaxesGroup FROM Indicators NATURAL JOIN Happiness15
      WHERE IndicatorCode='IC.TAX.TOTL.CP.ZS' GROUP BY TaxesGroup ORDER BY
      TaxesGroup;
```

Countries	MaxHappiness	TaxesGroup
4	6.611	10
18	7.527	20
33	7.587	30
39	7.522	40
23	7.364	50
10	7.226	60
7	6.983	70
2	5.89	80
1	6.574	140
1	3.956	220

6 Consultas do tipo relatório

De forma a atender o que foi solicitado na especificação como características avançadas, foi criada três consultas do tipo relatório que abrangem os aspectos mais gerais dos dados utilizados.

1. Felicidade e colocação dos países entre 2015 e 2017 em conjunto com o pontuação e ranking médio.

```
>> SELECT H1.CountryName AS Country, round(H1.HappinessScore,4) AS Score2015,
round(H2.HappinessScore,4) AS Score2016, round(H3.HappinessScore,4) AS
Score2017, round((((H1.HappinessScore + H2.HappinessScore) +
H3.HappinessScore) / 3),4) AS AvgScore, H1.HappinessRank AS Rank2015,
H2.HappinessRank AS Rank2016, H3.HappinessRank AS Rank2017,
round((((H1.HappinessRank + H2.HappinessRank) + H3.HappinessRank) / 3),2)
AS AvgRank from ((Happiness15 H1 join Happiness16 H2 on((H1.CountryName =
H2.CountryName))) join Happiness17 H3 on((H2.CountryName = H3.CountryName)));
```

Country	S2015	S2016	S2017	AvgScore	R2015	R2016	R2017	AvgRank
Afghanistan	3.5750	3.3600	3.7940	3.5763	153	154	141	149.33
Albania	4.9590	4.6550	4.6440	4.7527	95	109	109	104.33
Algeria	5.6050	6.3550	5.8720	5.9440	68	38	53	53.00
Angola	4.0330	3.8660	3.7950	3.8980	137	141	140	139.33
Argentina	6.5740	6.6500	6.5990	6.6077	30	26	24	26.67
Armenia	4.3500	4.3600	4.3760	4.3620	127	121	121	123.00
Australia	7.2840	7.3130	7.2840	7.2937	10	9	10	9.67
Austria	7.2000	7.1190	7.0060	7.1083	13	12	13	12.67
Azerbaijan	5.2120	5.2910	5.2340	5.2457	80	81	85	82.00
Bahrain	5.9600	6.2180	6.0870	6.0883	49	42	41	44.00
...

Vale atentar que essa *query* foi transformada em um View, nomeada como **HappinessInTime** de modo a permitir a sua utilização para outras queries de forma mais eficiente. Outro detalhe diz respeito ao nome de algumas colunas estarem diferentes da *query*. Isso é explicado pelo fato dessas terem sido abreviadas com o intuito de reduzir a largura da tabela e essa conseguir ser exibida corretamente nesse documento.

2. Média da colocação dos países considerando região e *IncomeGroup*.

```
>> SELECT AVG(HappinessRank) AvgRank, C.Region, IncomeGroup FROM Country C
JOIN Happiness15 H ON C.ShortName = H.CountryName GROUP BY C.Region,
IncomeGroup ORDER BY AvgRank DESC;
```

AvgRank	Region	IncomeGroup
145.0000	East Asia & Pacific	Low income
139.0000	Sub-Saharan Africa	Low income
137.0000	South Asia	Low income
122.2500	Middle East & North Africa	Lower middle income
119.0000	Latin America & Caribbean	Low income
118.4000	Sub-Saharan Africa	Upper middle income
111.7000	Sub-Saharan Africa	Lower middle income
103.6000	South Asia	Lower middle income
92.1429	Middle East & North Africa	Upper middle income
92.0000	East Asia & Pacific	Lower middle income
91.2857	Europe & Central Asia	Lower middle income
84.3333	Europe & Central Asia	Upper middle income
69.7500	East Asia & Pacific	Upper middle income
67.6000	Europe & Central Asia	High income: nonOECD
59.6000	Latin America & Caribbean	Lower middle income
42.0000	Latin America & Caribbean	Upper middle income
33.6087	Europe & Central Asia	High income: OECD
32.8571	Middle East & North Africa	High income: nonOECD
31.5000	Latin America & Caribbean	High income: nonOECD
27.0000	Latin America & Caribbean	High income: OECD
24.0000	East Asia & Pacific	High income: nonOECD
21.6667	East Asia & Pacific	High income: OECD
11.0000	Middle East & North Africa	High income: OECD
10.0000	North America	High income: OECD

3. O rank e a relação da família na felicidade, em conjunto com a mortalidade infantil por 1000 nascimentos, detalhada em neonatal, menos que um ano e menos que cinco anos.

```
>> SELECT HappinessRank, ROUND((Family/HappinessScore),4) HappyFamily,
      I2.Value '<4weeks', I3.Value '<1year', I4.Value '<5years' FROM
      (((Happiness15 H JOIN Indicators I2 ON H.CountryName = I2.CountryName)
      JOIN Indicators I3 ON I2.CountryName = I3.CountryName) JOIN Indicators I4
      ON I3.CountryName = I4.CountryName) WHERE I2.IndicatorCode='SP.DYN.IMRT.IN'
      AND I3.IndicatorCode='SH.DYN.NMRT' AND I4.IndicatorCode='SP.DYN.IMRT.IN'
      ORDER BY HappinessRank;
```

HappinessRank	HappyFamily	<4weeks	<1year	<5years
1	0.1779	3.4	2.7	3.4
2	0.1855	1.6	0.9	1.6
3	0.1807	2.9	2.5	2.9
4	0.1769	2	1.5	2
5	0.1781	4.3	3.2	4.3
6	0.1780	1.9	1.3	1.9
7	0.1735	3.2	2.4	3.2
8	0.1750	2.4	1.6	2.4
9	0.1811	4.7	3.1	4.7
10	0.1797	3	2.2	3
...

7 Consultas com dados não convencionais

Com o intuito de atender ao requisito que se referia as consultas envolvendo dados não convencionais como texto, imagem, som, vídeo dentro outros, foi acrescentado ao banco de dados, na tabela `Country` e no formato `blob`, as bandeiras dos países existentes. A latitude e longitude dos países também foi adicionada, o que possibilitou novas possibilidades de pesquisa como a que se segue.

O rank e a relação da família na felicidade, em conjunto com a mortalidade infantil por 1000 nascimentos, detalhada em neonatal, menos que um ano e menos que cinco anos.

```
>> SELECT HIT.AvgScore, ShortName, Flag, GeoLat, GeoLong FROM Country C JOIN
HappinessInTime HIT ON C.ShortName = HIT.Country WHERE GeoLat < 0 ORDER BY
HIT.AvgScore DESC;
```

AvgScore	ShortName	Flag	GeoLat	GeoLong
7.3113	New Zealand	[BLOB - 6.5 KiB]	-40.900557	147.885971
7.2937	Australia	[BLOB - 6.9 KiB]	-25.274398	133.775139
6.8567	Brazil	[BLOB - 15.8 KiB]	-14.235004	-51.92528
6.6757	Chile	[BLOB - 2.2 KiB]	-35.675147	-71.542969
6.6077	Argentina	[BLOB - 12 KiB]	-38.416097	-63.616672
6.4947	Uruguay	[BLOB - 16.1 KiB]	-32.522779	-55.765835

8 Interface

Outra característica presente no trabalho realizado é a presença de uma interface gráfica que permite ao usuário ver as consultas e seus resultados, assim como fazer a sua própria consulta por SQL.

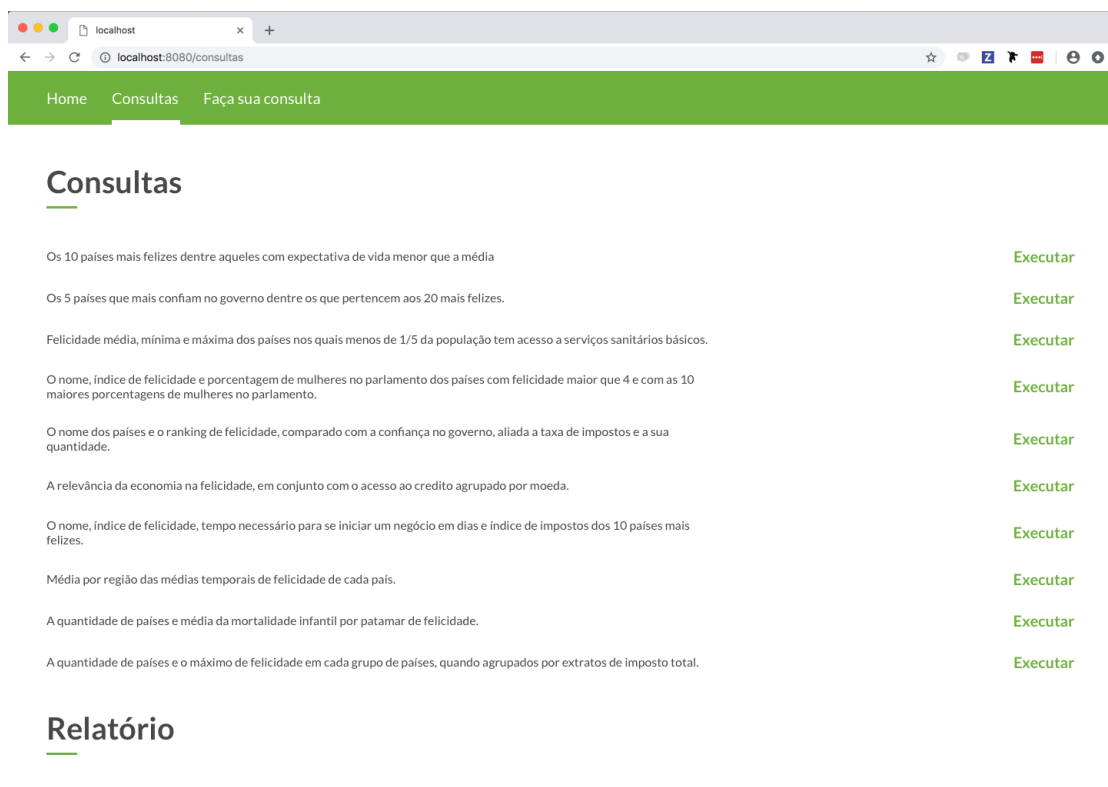


Figura 3 Página mostrando as consultas disponíveis.

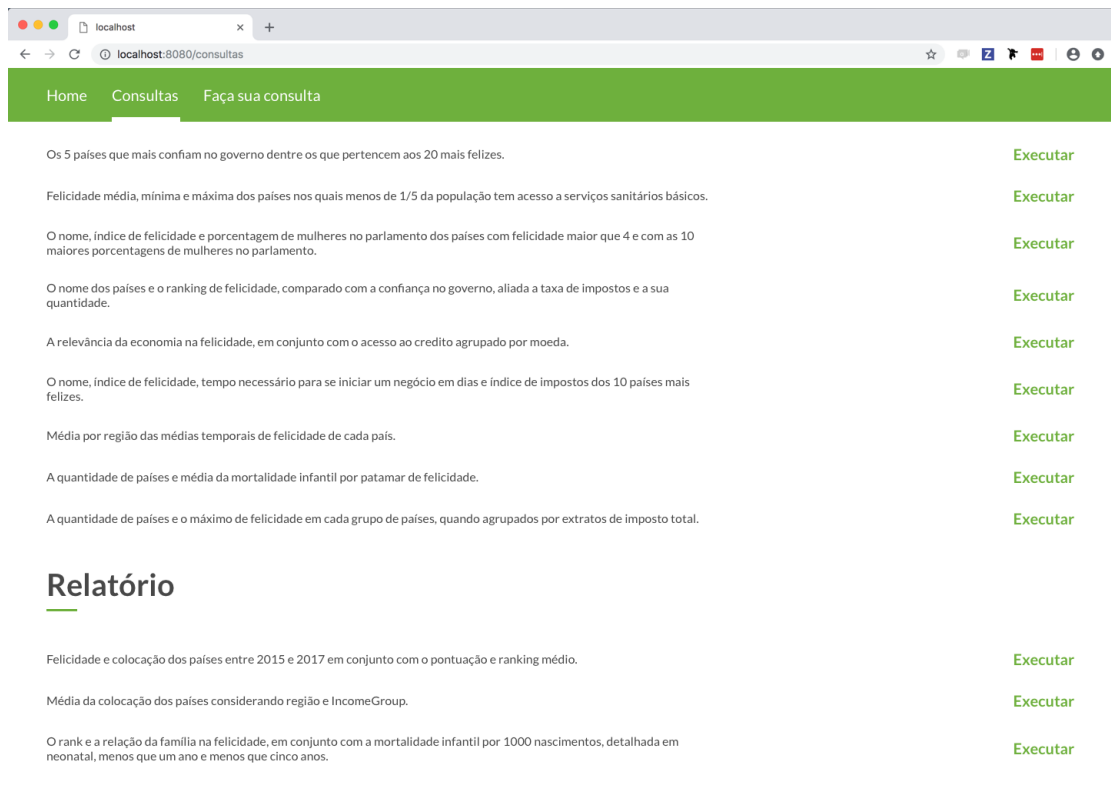


Figura 4 Página mostrando as consultas do tipo relatório disponíveis.

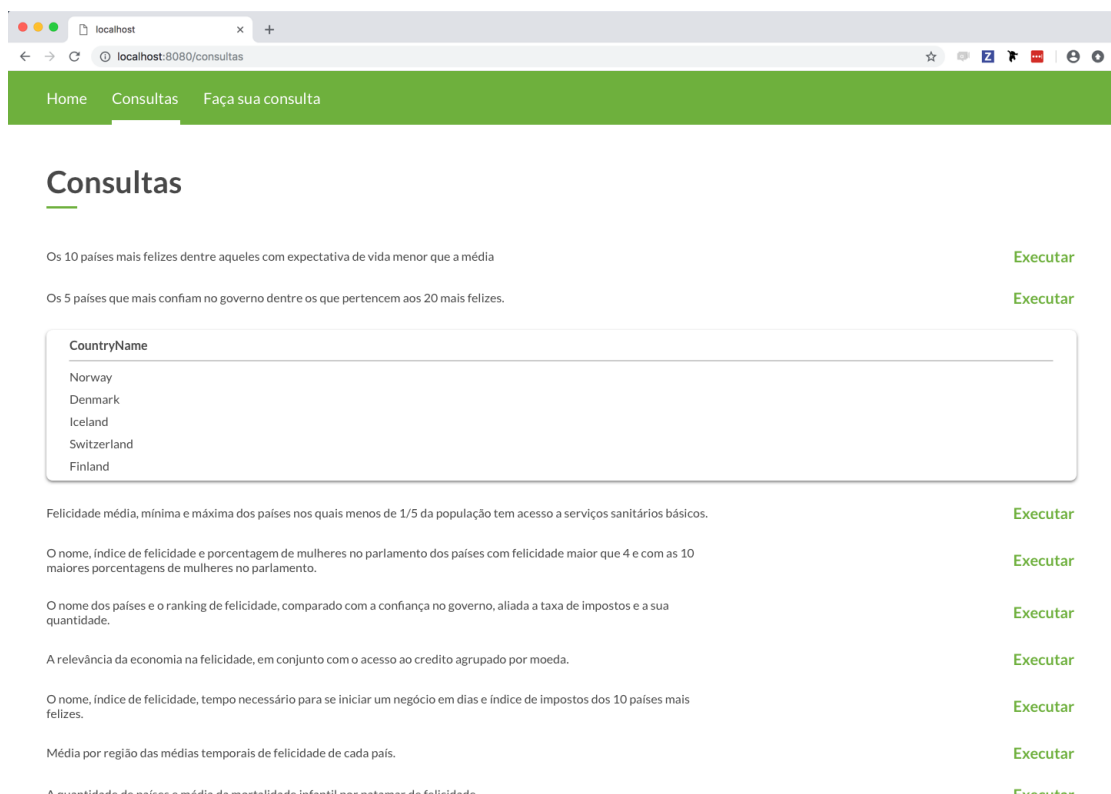


Figura 5 Página mostrando o resultado de uma das consultas disponíveis.

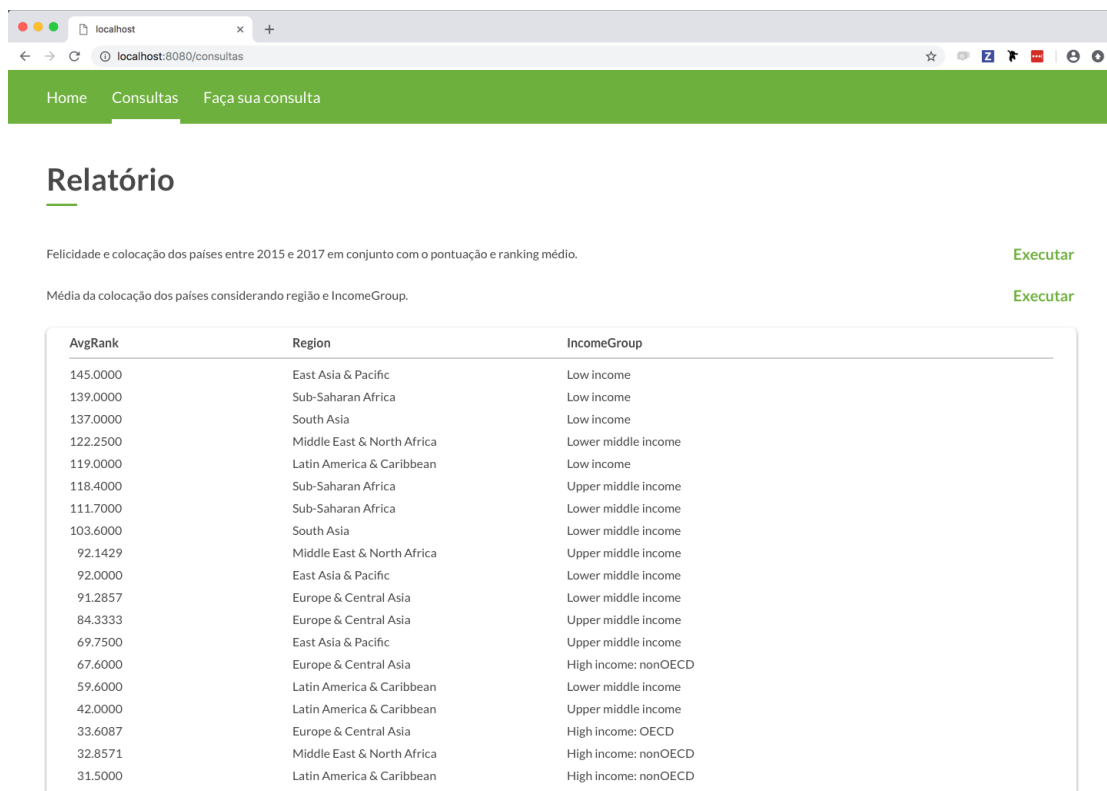


Figura 6 Página mostrando o resultado de uma das consultas do tipo relatório disponíveis.

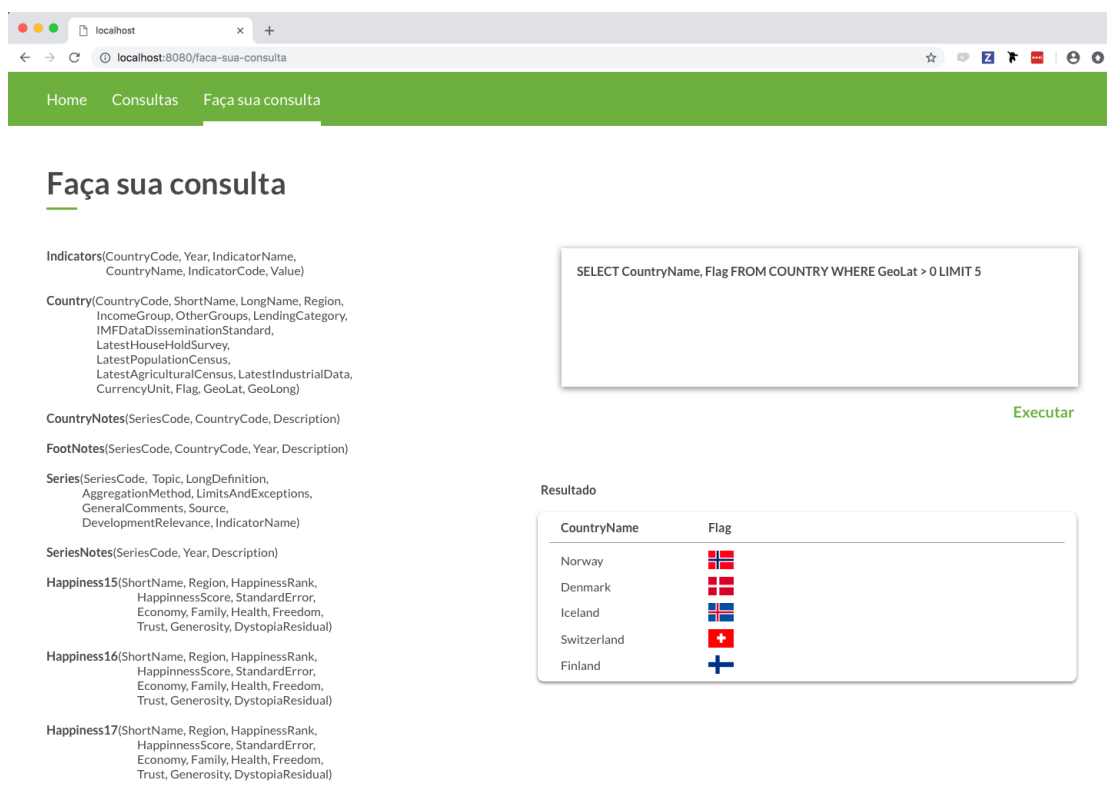


Figura 7 Página mostrando a interface para consultas personalizadas.

9 Atividades do grupo

De forma sucinta, a divisão de trabalho no grupo se deu da seguinte maneira:

- **Artur Marzano.** Construção de consultas alto-nível e criação/testes dos SQLs correspondentes. Coleta e importação dos dados não convencionais.
- **João Couto.** Modelagem dos datasets e desenvolvimento dos diagramas ER e relacional.
- **João Victor.** Escrita da parte textual da documentação e criação da interface.
- **Tiago Alves.** Modelagem, filtragem dos datasets, revisão da documentação e liderança do grupo.

10 Conclusão

Consideramos que o trabalho foi concluído com relativo sucesso. As atividades que mais demandaram tempo foram a filtragem e compreensão do formato dos dados e o processo de criação das consultas. O trabalho foi de suma importância para o desenvolvimento de habilidades a respeito da matéria ministrada na disciplina, pois, embora alguns dos membros já tivessem alguma experiência prévia com aplicações apoiadas em bancos de dados, a modelagem do zero, tanto do banco quanto das consultas, se mostrou uma prática interessante e educativa.

11 Referências Bibliográficas

Elmasri, Navathe. *Fundamentals of Database Systems*.

Peter Chen, *The Entity-Relationship Model Toward A Unified View of Data*

E. F. Codd, *A Relational Model of Data for Large Shared Data Banks*