



IPS  
Instituto  
Politécnico de Setúbal  
Escola Superior de  
Tecnologia do Barreiro

# Morbidity and Mortality



Bernardo Augusto & Miguel Cisneiros

Big Data

Professor Raquel Barreira

BSc in Bioinformatics – 2020/2021

# Overview of the task

**The goal:** To explore a document-oriented database in MongoDB



## Methodology

1. Check open-access database sources
2. Select a national database
3. Upload and work it on MongoDB (with queries and aggregations)



# The document dataset

- This dataset is related to the monthly evolution of hospitalization, outpatient and death episodes by main diagnosis category, between January 2016 to September 2018 at the Cascais Hospital - Dr. José de Almeida.
- It is provided in a JSON file (JavaScript Object Notation) and contains the following attributes:



- datasetid – identification of the dataset
- recordid – identification of the row
- fields/obitos – number of deaths occurred
- fields/dias\_internamento – number of days of hospitalization
- fields/faixa\_etaria – age range of the patient
- fields/periodo – month and year of the episode
- fields/desc\_capitulo – main diagnosis category
- fields/sexo – gender of the patient
- fields/regi\_es - health region of the hospital
- fields/internamentos – number of hospitalized patients
- fields/instituicao – name of the hospital
- fields/cod\_capitulo – code of the main diagnosis category
- fields/ambulatorio – number of outpatients
- record\_timestamp – record date and time

**cascais.cascais**

COLLECTION SIZE: 2.39MB

TOTAL DOCUMENTS: 5004

INDEXES TOTAL SIZE: 72KB

# Query 1

Q: What was the main cause of death in the timeframe of the dataset?

The screenshot shows the MongoDB aggregation pipeline interface. It consists of two main sections: the top section displays the output after the \$group stage, and the bottom section displays the output after the \$sort stage.

**Top Section: Output after \$group stage (Sample of 20 documents)**

Aggregation Pipeline Stage: \$group

```
1 /**
2 * _id: The id of the group.
3 * fieldN: The first field name.
4 */
5 {
6   id: "$fields.desc_capitulo",
7   total_obitos: {
8     $sum: "$fields.obitos"
9   }
10 }
```

Output (Sample of 20 documents):

- `_id: "Doenças do aparelho respiratório"`  
`total: 325`
- `_id: "Doenças do aparelho geniturinário"`  
`total: 85`

**Bottom Section: Output after \$sort stage (Sample of 20 documents)**

Aggregation Pipeline Stage: \$sort

```
1 /**
2 * Provide any number of field/order pairs.
3 */
4 {
5   total_obitos: -1
6 }
```

Output (Sample of 20 documents):

- `_id: "Doenças do aparelho respiratório"`  
`total_obitos: 325`
- `_id: "Neoplasias"`  
`total_obitos: 306`

A blue oval highlights the document from the \$group stage output, which corresponds to the document highlighted in the \$sort stage output.

A: Diseases Of The Respiratory System

# Query 2

Q: What was the month with more deaths?

The screenshot shows the MongoDB aggregation pipeline interface with two stages displayed.

**\$group Stage:**

- Buttons: \$group, green toggle switch, trash, plus.
- Output title: "Output after \$group stage" (Sample of 20 documents).
- Code:

```
1 /**
2  * _id: The id of the group.
3  * fieldN: The first field name.
4 */
5 {
6   _id: "$fields.periodo",
7   total_obitos: {
8     $sum: "$fields.obitos"
9   }
10 }
```
- Output documents (Sample of 20):
  - `_id: "2018-06" total_obitos: 46`
  - `_id: "2018-07" total_obitos: 43`
  - `_id: "2017-10" total_obitos: 53`

**\$sort Stage:**

- Buttons: \$sort, green toggle switch, trash, plus.
- Output title: "Output after \$sort stage" (Sample of 20 documents).
- Code:

```
1 /**
2  * Provide any number of field/order pairs.
3 */
4 {
5   total_obitos: -1
6 }
```
- Output documents (Sample of 20):
  - `_id: "2016-12" total_obitos: 76` (This document is circled in blue.)
  - `_id: "2017-12" total_obitos: 75`
  - `total_obitos: 71 _id: "2017-02"`

A: December 2016 with a total of 76 deaths

# Query 3

Q: What was the age group with more deaths in the timeframe of the dataset?

The screenshot shows the MongoDB aggregation pipeline interface with two stages displayed.

**\$group Stage:**

- Stage name: \$group
- Enabled: Yes
- Output after \$group stage (Sample of 7 documents):
- Document 1: total\_obitos: 4, \_id: "[15-25["
- Document 2: \_id: "[65-120[", total\_obitos: 1631
- Document 3: \_id: "[0-1[", total\_obitos: 5

**\$sort Stage:**

- Stage name: \$sort
- Enabled: Yes
- Output after \$sort stage (Sample of 7 documents):
- Document 1: \_id: "[65-120[", total\_obitos: 1631 (circled in blue)
- Document 2: \_id: "[45-65[", total\_obitos: 187
- Document 3: \_id: "[25-45[", total\_obitos: 20

A: People from 65 to 120 years old were the ones with more deaths

# Query 4

Q: In average, what was the age group with more days of hospitalization?

The screenshot shows the MongoDB aggregation pipeline interface. It consists of two main sections: the top section displays the output after the \$group stage, and the bottom section displays the output after the \$sort stage.

**Output after \$group stage:**

```
1 /**
2  * _id: The id of the group.
3  * fieldN: The first field name.
4  */
5 {
6   _id: "$fields.faixa_etaria",
7   media_dias_internamento: {
8     $avg: "$fields.dias_internamento"
9   }
10 }
```

**Output after \$sort stage:**

```
1 /**
2  * Provide any number of field/order pairs.
3  */
4 {
5   media_dias_internamento: -1
6 }
```

In the \$group stage output, three documents are shown, each representing an age group range and its average hospitalization days:

- `_id: "[0-1[" media_dias_internamento: 98.94811320754717`
- `_id: "[1-5[" media_dias_internamento: 8.483790523690773`
- `_id: "[5-15[" media_dias_internamento: 5.175767918088737`

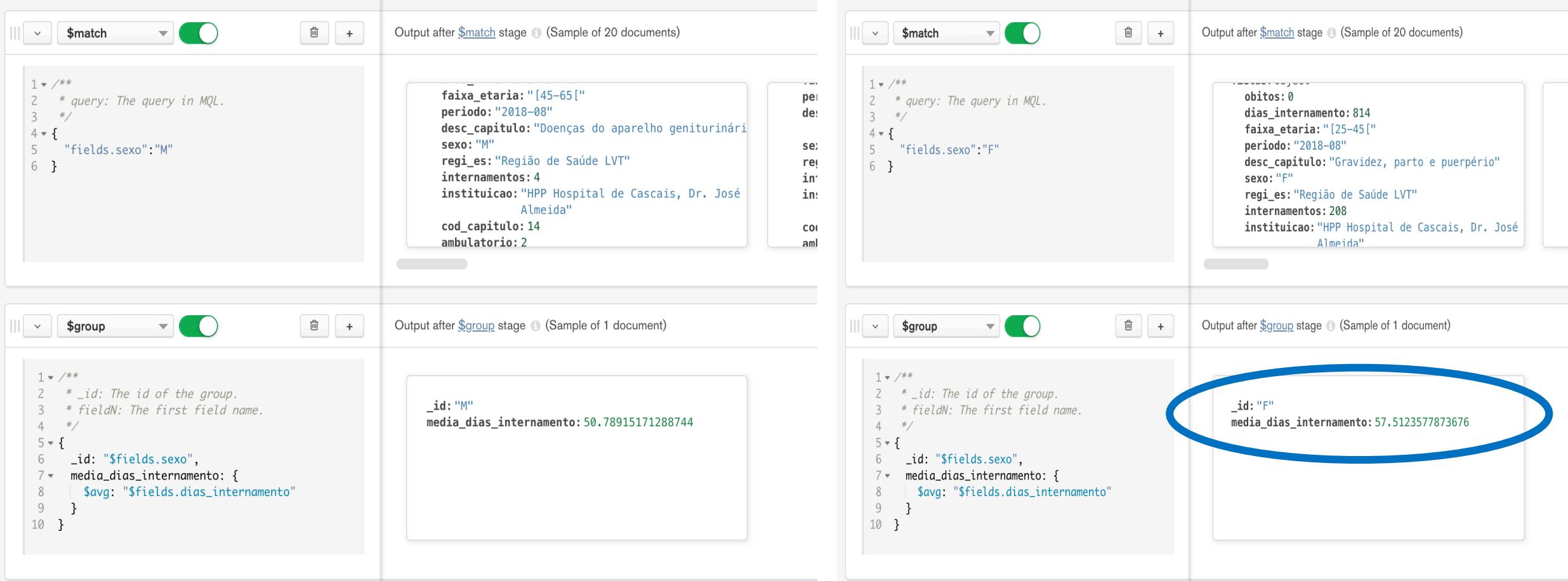
In the \$sort stage output, the documents are sorted by average hospitalization days in descending order. The first document, representing the age group [65-120[, is circled in blue:

- `_id: "[65-120[" media_dias_internamento: 128.87922705314008`
- `_id: "[0-1[" media_dias_internamento: 98.94811320754717`
- `_id: "[25-45[" media_dias_internamento: 43.809734513274336`

A: People from 65 to 120 years old were the ones with more days of hospitalization

# Query 5

**Q:** In average, what was the gender with more days of hospitalization?



The image shows two separate MongoDB aggregation pipelines. Each pipeline consists of a '\$match' stage followed by a '\$group' stage.

**Left Pipeline (\$match Stage):**

```

1 /**
2 * query: The query in MQL.
3 */
4 {
5   "fields.sexo": "M"
6 }
  
```

**Output after \$match stage (Sample of 20 documents):**

```

faixa_etaria: "[45-65["
periodo: "2018-08"
desc_capitulo: "Doenças do aparelho geniturinário"
sexo: "M"
regi_es: "Região de Saúde LVT"
internamentos: 4
instituicao: "HPP Hospital de Cascais, Dr. José Almeida"
cod_capitulo: 14
ambulatorio: 2
  
```

**Left Pipeline (\$group Stage):**

```

1 /**
2 * _id: The id of the group.
3 * fieldN: The first field name.
4 */
5 {
6   _id: "$fields.sexo",
7   media_dias_internamento: {
8     $avg: "$fields.dias_internamento"
9   }
10 }
  
```

**Output after \$group stage (Sample of 1 document):**

```

_id: "M"
media_dias_internamento: 50.78915171288744
  
```

**Right Pipeline (\$match Stage):**

```

1 /**
2 * query: The query in MQL.
3 */
4 {
5   "fields.sexo": "F"
6 }
  
```

**Output after \$match stage (Sample of 20 documents):**

```

obitos: 0
dias_internamento: 814
faixa_etaria: "[25-45["
periodo: "2018-08"
desc_capitulo: "Gravidez, parto e puerpério"
sexo: "F"
regi_es: "Região de Saúde LVT"
internamentos: 208
instituicao: "HPP Hospital de Cascais, Dr. José Almeida"
  
```

**Right Pipeline (\$group Stage):**

```

1 /**
2 * _id: The id of the group.
3 * fieldN: The first field name.
4 */
5 {
6   _id: "$fields.sexo",
7   media_dias_internamento: {
8     $avg: "$fields.dias_internamento"
9   }
10 }
  
```

**Output after \$group stage (Sample of 1 document):**

```

_id: "F"
media_dias_internamento: 57.5123577873676
  
```

**A:** Women had, in average, more days of hospitalization than men

# References

- MongoDB. (n.d.). *Get started with MongoDB — MongoDB Documentation.* MongoDB | Documentation.  
<https://docs.mongodb.com>
- MongoDB. (2019, June 13). *MongoDB Atlas - Aggregation Pipeline Builder Tutorial* [Video]. YouTube.  
[https://www.youtube.com/watch?v=0MZF\\_TiKIPnU](https://www.youtube.com/watch?v=0MZF_TiKIPnU)
- Barreira, R. (2020). *06 - MongoDB* [Slides]. Moodle | IPS.  
[https://moodle.ipb.pt/2021/pluginfile.php/197606/mod\\_resource/content/1/BD2021\\_06.pdf](https://moodle.ipb.pt/2021/pluginfile.php/197606/mod_resource/content/1/BD2021_06.pdf)
- Database  
<https://data.cascais.pt/pt-pt/node/254>



# Thank you

