Conteúdo

Professor Cipriano
Universidade Lusófona

May 19, 2022

Re: *Universidade do Pais das Maravilhas*

Dear Professor Cipriano,

I have completed the first phase of the analysis that you requested regarding the above referenced client as reported below. Please let me know if I can do anything further.

I. Summary

After analyzing the file, I knew immediately that I should create a new database in PhpMyadmin. I named the new database Universidade_Maravilhas and created a table within Universidade_Maravilhas named Aluno with seven columns matching the names of the SQL file. I choose "nr_aluno" as the primary key after learning that each such number was unique.[1] After this, I imported the file in to PhpMyadmin. Please ***See*** **Exhibit A("Photographs of PHP Admin") attached hereto.**

A. Issues Seen

Once the data was imported, I analyze it and noted the following:

1. The database appeared that it was not normalized. For example (as noted in footnote 1), there were several students with the same name. It was hard for me to imagine that so many students would have the same first and last name. Assuming that the names did not correspond to unique individuals, I imagined that we could create an alternative table to solve this problem with a sequential order corresponding to each unique individual. Before attempting to do this, I sent my inquiry to Professor Cipriano and was told that each person in the database was in fact a unique individual. In addition, I could have made a second chart for the courses as the values in this column contained repeat values. However, it did not seem to contribute to a more efficient way to analyze the data and therefore I did not create such a table.

[1] It was initially unclear if the same named students were distinct individuals. If the same named people were in part duplicates, that would have posed a problem to using the student number as the primary key. However, I confirmed that each student number corresponded to a unique person.

2. The table contained several entries of negative and/or null values in the columns “medial_entrade” and “media_final” which presented a problem when computing statistics such as the mean. In addition, I found that some of the entries in the column “curso” contained nonsensical responses. To handle these issues I applied the following filter:

```
//setp 1: write  down the query
$query = "SELECT * from aluno where media_entrada IS NOT NULL AND media_entrada > 0 AND
media_final IS NOT NULL AND media_final > 0 and  curso in ('Informática','Comunicação','Psicologia')
AND curso_concluido = 1
 order by nr_aluno ASC";
```

After cleaning up the data as set forth above, I extracted relevant data into a CSV file by using php code. See Exhibit B(“Php Code for CSV file and Phase1Dados.csv”). I also sorted the flawed data into a text file. ***See*** **Exhibit C (“Faultydados.txt”) attached hereto**. I converted this txt file to a CSV file so that I could include it into my exploratory analysis.

With all the data in order, I created a jupyter notebook name Universidade Maravilhas, imported the CSV files and ran some exploratory data, which included determining the range for the grades entering and exiting the program. When determining the maximum scored in both the entering grades and exiting grades, I noticed that these values were 21 and 22. I was not sure if I should filter these scores out as part of my faulty data as I believe they are not possible out of a score of 20. However, I did not filter them out because in the US we can achieve such scores. ***See*** **Exhibit D (“Jupyter Notebook”) attached hereto.**

II. Questions of Inquiry

1. Does there exist a relationship between the entering grades and exiting grades?
2. Does there exist a relationship between the student’s major and the exiting grade?

III. Conclusion/Results

By organizing and analyzing the data, I was able to respond to the questions of inquiry. Namely, 1) based on the cleaned data, there does not appear to be a correlation between the entering and exiting grades as the mean values of both are nearly equal, and 2) based on the clean data, there does not appear to be a correlation between the student’s entering major and the final grades because the mean with respect to the ending grades were substantially the same across the board. Please ***See*** **Exhibit E.**

EXHIBIT A

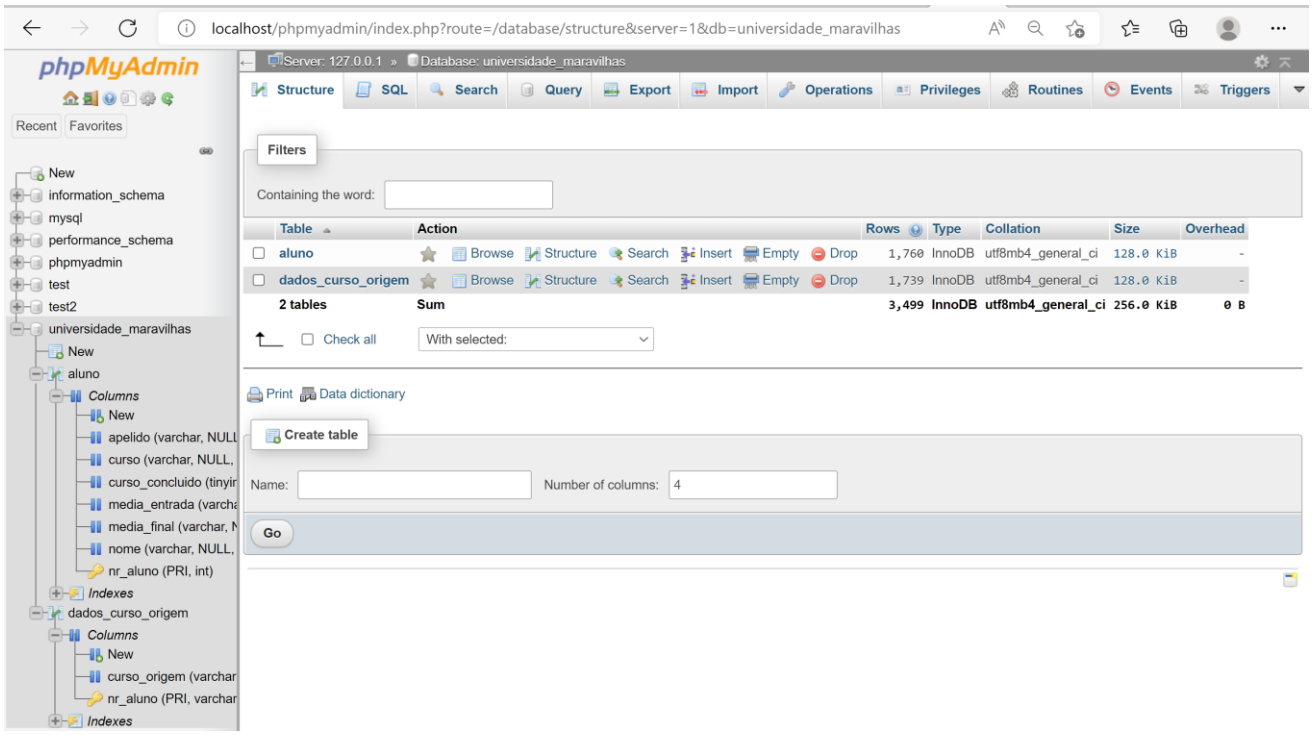

localhost/phpmyadmin/index.php?route=/database/structure&server=1&db=universidade_maravilhas
phpMyAdmin
Recent
Favorites
New
information_schema
mysql
performance_schema
phpmyadmin
test
test2
universidade_maravilhas
New
aluno
Columns
New
apelido (varchar, NULL
curso (varchar, NULL,
curso_concluido (tinyir
media_entrada (varcha
media_final (varchar, N
nome (varchar, NULL,
nr_aluno (PRI, int)
Indexes
dados_curso_origem
Columns
New
curso_origem (varchar
nr_aluno (PRI, varchar
Indexes
Server: 127.0.0.1 » Database: universidade_maravilhas
Structure
SQL
Search
Query
Export
Import
Operations
Privileges
Routines
Events
Triggers
Filters
Containing the word:
Table
Action
Rows
Type
Collation
Size
Overhead
aluno
Browse
Structure
Search
Insert
Empty
Drop
1,760
InnoDB
utf8mb4_general_ci
128.0 KiB
dados_curso_origem
Browse
Structure
Search
Insert
Empty
Drop
1,739
InnoDB
utf8mb4_general_ci
128.0 KiB
2 tables
Sum
3,499
InnoDB
utf8mb4_general_ci
256.0 KiB
0 B
Check all
With selected:
Print
Data dictionary
Create table
Name:
Number of columns:
4
Go

EXHIBIT B

```
<?php

$hostName = "localhost";
$username = "root";
$password = "";
$database = 'universidade_maravilhas';

$connection = new mysqli($hostName, $username, $password, $database);
// Create connection

// Check connection
if ($connection->connect_error) {
  die("Connection failed: " . $conn->connect_error);
}

//setp 1: write  down the query
$query = "SELECT * from aluno where media_entrada IS NOT NULL AND
media_entrada > 0 AND
 media_final IS NOT NULL AND media_final > 0 and  curso in
('Informática','Comunicação','Psicologia')
 AND curso_concluido = 1
  order by nr_aluno ASC";

//setp2: execute query
$execute = mysqli_query($connection, $query);



$delimiter = ",";
// $filename = "name". ".csv";
$filename = 'Phase1dados.csv';

// Create a file pointer
//create a temporary file in the memory
//allow it to write

$f = fopen('php://memory', 'w');

// Set column headers
$fields = array('nr_aluno', 'nome', 'apelido', 'curso', 'media_entrada',
'curso_concluido', 'media_final');

fputcsv($f, $fields, $delimiter);


//step 3: fetch the records using a loop

while($row = mysqli_fetch_array($execute)){
```

```
    $nr_aluno    = $row['nr_aluno'];
    $nome        = $row['nome'];
    $apelido     = $row['apelido'];
    $curso       =  $row['curso'];
    $media_entrada = $row['media_entrada'];
    $curso_concluido = $row['curso_concluido'];
    $media_final     =   $row['media_final'];


    $rowArray = array($nr_aluno, $nome, $apelido, $curso, $media_entrada,
$curso_concluido, $media_final);

    fputcsv($f, $rowArray, $delimiter);

 }

  // Move back to beginning of file
  fseek($f, 0);

  // Set headers to download file rather than displayed
  header('Content-Type: text/csv');
  header('Content-Disposition: attachment; filename="' . $filename . '";');

  //output all remaining data on a file pointer
  fpassthru($f);
```

ır_aluno	nome	apelido	curso	media_entrada	curso_concluido	media_final
20216	Joana	Castelo	Comunicação	20	1	12
200315	Anne	Morais	Psicologia	14	1	16
200321	Pedro	Castelo	Informática	10	1	19
200326	Patrícia	da Silva	Psicologia	20	1	10
200331	Beatriz	Castelo	Comunicação	19	1	12
200332	Bruno	Machado	Comunicação	18	1	10
200333	Alice	Sampaio	Psicologia	11	1	17
200335	Diana	Morais	Psicologia	14	1	15
200346	Olivia	Cintra	Comunicação	19	1	12
200355	Raquel	Cintra	Comunicação	12	1	17
200356	Anne	Machado	Psicologia	16	1	16
200362	Rita	Castelo-Branco	Comunicação	17	1	12
200367	Anne	Machado	Psicologia	14	1	16
200370	Olivia	Silva	Psicologia	20	1	12
200371	Jorge	Laranjeira	Psicologia	18	1	11
200374	Sandra	da Silva	Comunicação	11	1	20
200380	Nuno	Silva	Comunicação	10	1	20
200381	Rita	Cintra	Informática	18	1	12
200389	Joana	da Silva	Psicologia	10	1	19
200391	Carla	Laranjeira	Psicologia	13	1	20
200397	Miguel	Castelo-Branco	Comunicação	16	1	15
200714	Rui	Sampaio	Comunicação	17	1	12
200719	Maria João	Pereira	Psicologia	10	1	19
200728	Maria	Silva	Informática	12	1	20
200729	Rodrigo	Cintra	Informática	10	1	20
200734	Ana	Morais	Psicologia	14	1	14
200737	Lucas	Capicua	Psicologia	18	1	11
200755	Raquel	Castelo	Comunicação	14	1	16
200757	Maria João	Capicua	Psicologia	10	1	17

alunos_6 | Folha1

- **Please note that I have attached this file in GitHub for your review.**

EXHIBIT C

nr_aluno	nome	apelido	curso	media_entrada	curso_concluido	media_final
200722	Catarina	da Silva	Psicologia	13	0	
200739	Alice	Pires	Psicologia	17	0	
202245	Rui	Pires	Informática	10	1	-2
2003485	Alice	Pires	Psicologia	13	0	
2003504	Beatriz	Laranjeira	Comunicação	12	0	
2003821	Nuno	Laranjeira	Psicologia	16	0	
2007103	Catarina	Castelo	Informática	14	0	
2007469	Nuno	da Silva	Psicologia	14	1	-1
2007691	Miguel	Silva	Informática	14	1	-2
2019514	Diana	Sampaio	Informática	-2	1	12
2020188	Nuno	Cintra	Desconhecido	15	1	16
2020254	Michaelangelo	Morais	Medicina Veterinária	12	1	20
2020264	Miguel	Pires	Medicina Veterinária	13	1	20
2020537	Maria João	Sampaio	Comunicação	12	0	
2021320	Jorge	Pires	Medicina Veterinária	17	1	11
2021715	Avelino	da Silva	Desconhecido	16	1	15
2021847	Patrícia	Pires	Psicologia	10	0	
2022848	Carla	Pereira	Psicologia	17	0	
2210030	Sininho	(Terra do Nunca)	Cintilar	-20	1	-20
2300910	Alf	Alien Life Form	Má Vida	20	1	20
20031072	Nuno	Pires	Comunicação	14	0	
20071679	Jorge	Sampaio	Comunicação	10	1	-1
20191494	Joana	Castelo	Psicologia	15	0	
20191750	Anne	Cintra	Comunicação	12	0	
20201046	Miguel	Silva	Psicologia	-1	1	17
20201546	Anne	Sampaio	Informática	15	1	
20211052	Raquel	Castelo	Comunicação	16	0	
20211156	Ana	Morais	Comunicação	20	1	
20211328	Ana	Sampaio	Informática	12	1	-1
20211394	Bruno	Cintra	Medicina Veterinária	18	1	13
20211664	Beatriz	Capicua	Informática	16	0	
20221087	Rodrigo	Morais	Informática	16	0	
20221456	Lucas	Silva	Informática	18	1	-1

EXHIBIT D

jupyter Universidade_Maravilhas (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Run Code

```
In [70]: import pandas as pd

In [71]: import matplotlib.pyplot as plt

In [72]: dados = pd.read_csv('C:/Users/faria/OneDrive/Desktop/python/Phase1dados.csv')

In [73]: dados.head() #initially the data was sep by semicolon which produced flawed date
         #realized that it was not being delinated by comma,changed the error.  working fine now
```

Out[73]:

	nr_aluno	nome	apelido	curso	media_entrada	curso_concluido	media_final
0	20216	Joana	Castelo	Comunicação	20	1	12
1	200315	Anne	Morais	Psicologia	14	1	16
2	200321	Pedro	Castelo	Informática	10	1	19
3	200326	Patrícia	da Silva	Psicologia	20	1	10
4	200331	Beatriz	Castelo	Comunicação	19	1	12

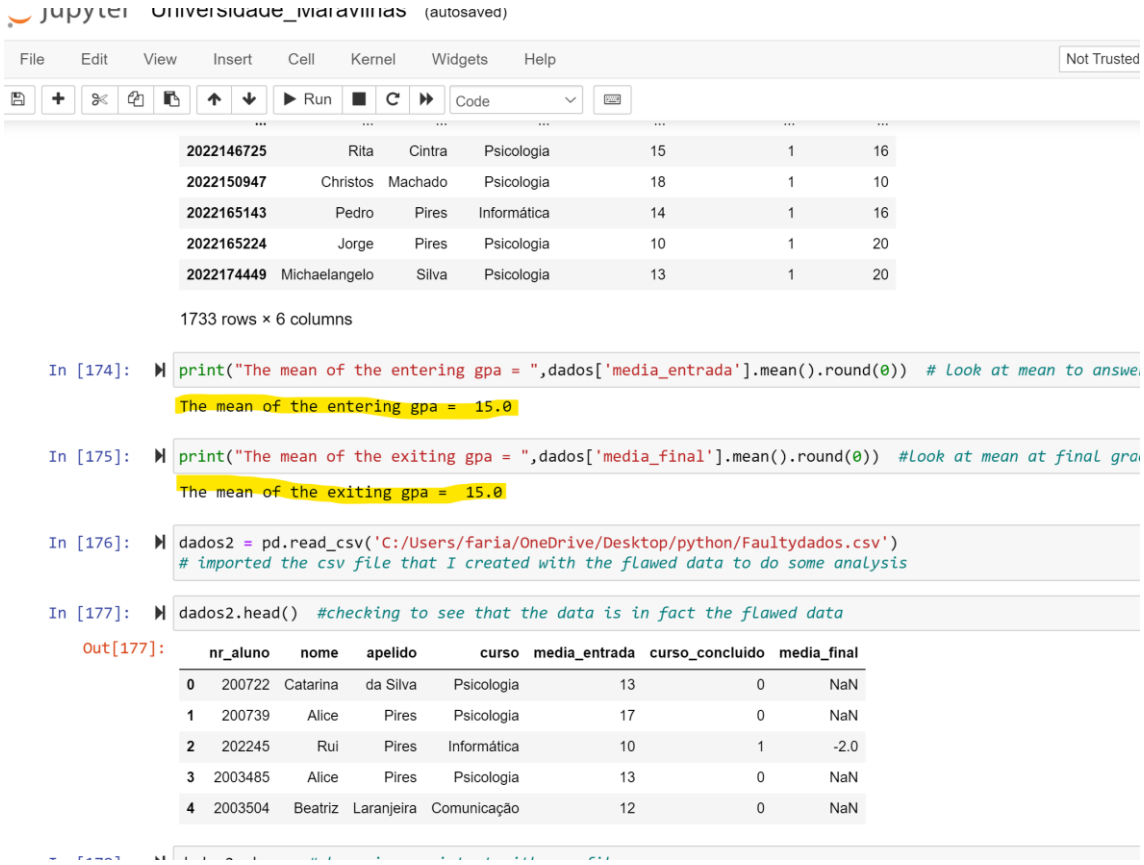
```
In [74]: dados['media_entrada']
```

Out[74]:

```
0       20
1       14
2       10
3       20
4       19
        ..
1717    15
1718    18
1719    14
1720    10
1721    13
Name: media_entrada, Length: 1722, dtype: int64
```

Please note that this notebook has been provided in Github

EXHIBIT E



jupyter Universidade_Maravilhas (autosaved)

File Edit View Insert Cell Kernel Widgets Help Not Trusted

Run Code

2022146725	Rita	Cintra	Psicologia	15	1	16
2022150947	Christos	Machado	Psicologia	18	1	10
2022165143	Pedro	Pires	Informática	14	1	16
2022165224	Jorge	Pires	Psicologia	10	1	20
2022174449	Michaelangelo	Silva	Psicologia	13	1	20

1733 rows × 6 columns

In [174]:
```
print("The mean of the entering gpa = ",dados['media_entrada'].mean().round(0))  # look at mean to answe
```
```
The mean of the entering gpa =  15.0
```

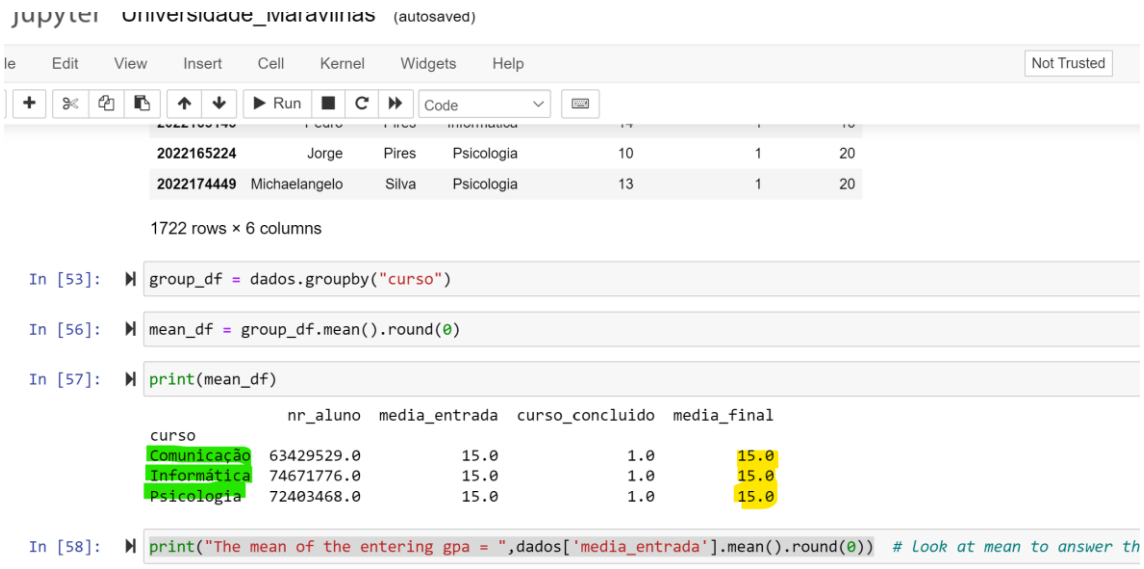
In [175]:
```
print("The mean of the exiting gpa = ",dados['media_final'].mean().round(0))  #Look at mean at final gra
```
```
The mean of the exiting gpa =  15.0
```

In [176]:
```
dados2 = pd.read_csv('C:/Users/faria/OneDrive/Desktop/python/Faultydados.csv')
# imported the csv file that I created with the flawed data to do some analysis
```

In [177]:
```
dados2.head()  #checking to see that the data is in fact the flawed data
```

Out[177]:

	nr_aluno	nome	apelido	curso	media_entrada	curso_concluido	media_final
0	200722	Catarina	da Silva	Psicologia	13	0	NaN
1	200739	Alice	Pires	Psicologia	17	0	NaN
2	202245	Rui	Pires	Informática	10	1	-2.0
3	2003485	Alice	Pires	Psicologia	13	0	NaN
4	2003504	Beatriz	Laranjeira	Comunicação	12	0	NaN



jupyter Universidade_Maravilhas (autosaved)

le Edit View Insert Cell Kernel Widgets Help Not Trusted

Run Code

2022165224	Jorge	Pires	Psicologia	10	1	20
2022174449	Michaelangelo	Silva	Psicologia	13	1	20

1722 rows × 6 columns

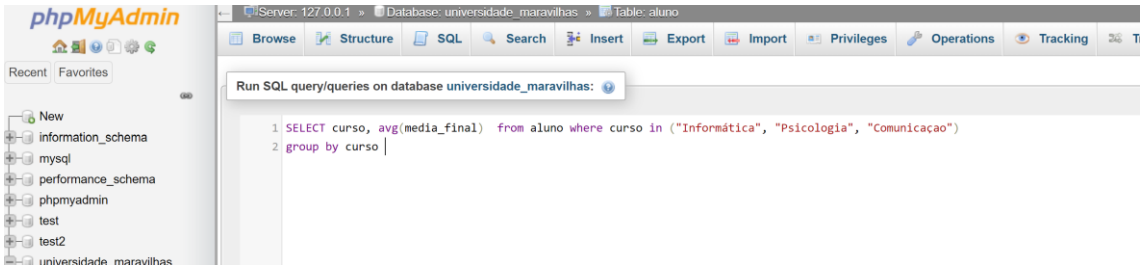
In [53]:
```
group_df = dados.groupby("curso")
```

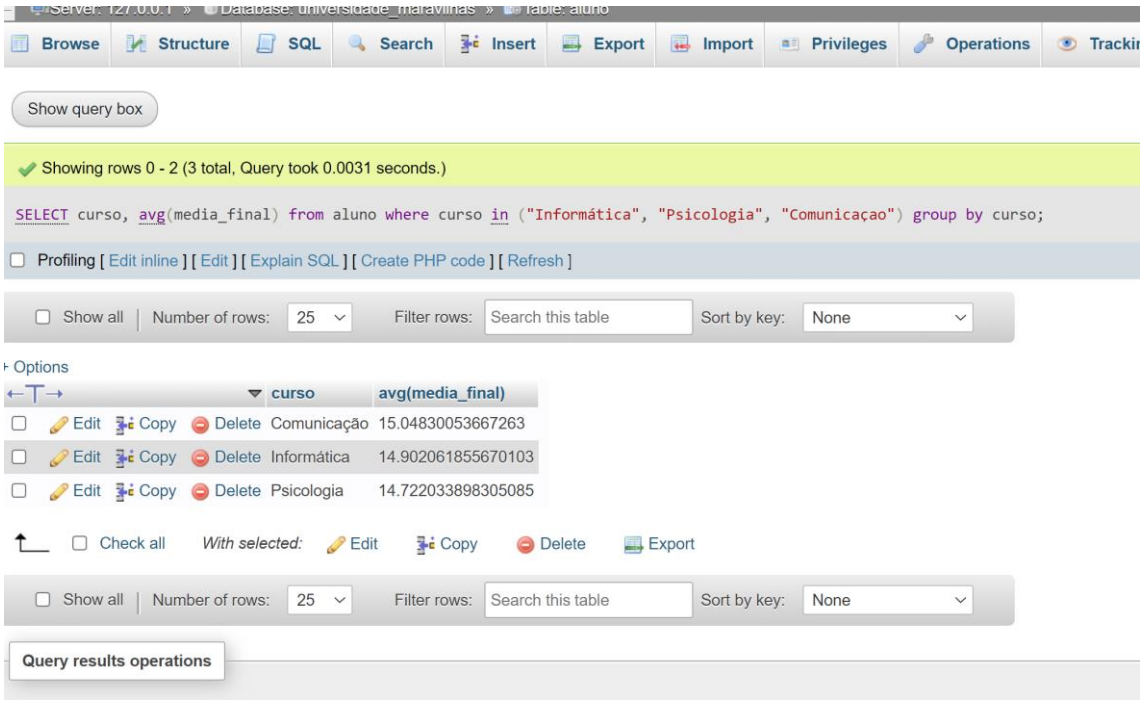
In [56]:
```
mean_df = group_df.mean().round(0)
```

In [57]:
```
print(mean_df)
```
```
                nr_aluno  media_entrada  curso_concluido  media_final
curso
Comunicação  63429529.0           15.0              1.0         15.0
Informática  74671776.0           15.0              1.0         15.0
Psicologia   72403468.0           15.0              1.0         15.0
```

In [58]:
```
print("The mean of the entering gpa = ",dados['media_entrada'].mean().round(0))  # look at mean to answer th
```

To be sure, I ran a sql query to confirm the results


phpMyAdmin
Recent
Favorites
New
information_schema
mysql
performance_schema
phpmyadmin
test
test2
Server: 127.0.0.1 » Database: universidade_maravilhas » Table: aluno
Browse
Structure
SQL
Search
Insert
Export
Import
Privileges
Operations
Tracking
Run SQL query/queries on database universidade_maravilhas:
SELECT curso, avg(media_final) from aluno where curso in ("Informática", "Psicologia", "Comunicaçao")
group by curso


Browse
Structure
SQL
Search
Insert
Export
Import
Privileges
Operations
Show query box
Showing rows 0 - 2 (3 total, Query took 0.0031 seconds.)
SELECT curso, avg(media_final) from aluno where curso in ("Informática", "Psicologia", "Comunicaçao") group by curso;
Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]
Show all
Number of rows: 25
Filter rows: Search this table
Sort by key: None
Options
curso
avg(media_final)
Edit Copy Delete Comunicação 15.04830053667263
Edit Copy Delete Informática 14.902061855670103
Edit Copy Delete Psicologia 14.722033898305085
Check all
With selected:
Edit
Copy
Delete
Export
Query results operations