

目錄

Prerequisite

1. Please write code in Colab to read the table and report its size (# of columns and rows)

2. Which row looks like an anomaly and should be removed? Why?

3. Are there more girls than boys?

4. Which college has the most graduates?

5. Which college has the most unbalanced 入學管道? In other words, we say it's balanced if each 管道 has a similar number of graduates.

Reference

Prerequisite

I use *PyDrive* to load the data from Google Drive. In order to use tables, I also import all of the modules called *datascience*. And, we need to do visualization, so I import *matplotlib*.

```
Prerequisite

Using PyDrive to load the data from Google Drive.

We are first required to install the PyDrive library f from the python installer(pip) and execute the following.

References from Ways to import CSV files in Google Colab - GeekforGeeks.

[170] # Code to read csv file into Colaboratory:
!pip install -U -q PyDrive
from pydrive.auth import GoogleAuth
from pydrive.drive import GoogleDrive
from google.colab import auth
from oauth2client.client import GoogleCredentials
# Authenticate and create the PyDrive client.
auth.authenticate_user()
gauth = GoogleAuth()
gauth.credentials = GoogleCredentials.get_application_default()
drive = GoogleDrive(gauth)

Import library we needed.

[180] import numpy as np
from datascience import *

import matplotlib
%matplotlib inline
import matplotlib.pyplot as plots
plots.style.use('fivethirtyeight')
```

1. Please write code in Colab to read the table and report its size (# of columns and rows)

I use the Table method *read_table* to read our imported CSV file and use *show()* method to print out the entire table to find the anomaly row.

1. Please write code in Colab to read the table and report its size (# of columns and rows)

```
[7] # The sharable link of data we would use
link = 'https://drive.google.com/file/d/1-lpsROWH91P54OGDSakfCfkCuEgQV1lr/view?usp=sharing'

# to get the id part of the file
id = link.split('/')[-2]

file_name = 'group_summary_table_20220929.csv'

downloaded = drive.CreateFile({'id':id})
downloaded.GetContentFile(file_name)

tbl = Table.read_table(file_name)
tbl.show() # show the entire table to find the anomaly row
```

Code Part

性別	入學管道	離校院別名稱(中)	num_students	credits_mean	credits_std	gpa_mean	gpa_std
女	學士班(指考分發/聯考)	地球科學學院	95	152.642	15.29	78.1504	7.17845
女	學士班(指考分發/聯考)	客家學院	17	151.824	16.7639	80.1531	6.12428
女	學士班(指考分發/聯考)	工學院	174	151.075	16.5397	78.8585	6.41941
女	學士班(指考分發/聯考)	文學院	686	154.219	20.8801	79.4352	5.83358
女	學士班(指考分發/聯考)	文學院中國文學系	1	172	nan	85.0756	nan
女	學士班(指考分發/聯考)	理學院	191	149.953	13.3548	77.0528	6.97312
女	學士班(指考分發/聯考)	生醫理工學院	76	151.961	17.8044	78.398	7.0676
女	學士班(指考分發/聯考)	管理學院	756	152.718	14.3223	78.5168	6.68581
女	學士班(指考分發/聯考)	資訊電機學院	119	150.126	14.8715	81.3947	6.56094

Output Table

Then, I use the *num_rows* and *num_columns* methods in the Table object type to print out its size. We can see that the **# of columns is 8** and the **# of rows is 49**.

Print out the # of rows and the # of this table.

```
[7] print("# of columns:", tbl.num_columns)
    print("# of rows: ", tbl.num_rows)
```

Code Part

```
# of columns: 8
# of rows: 49
```

Output

2. Which row looks like an anomaly and should be removed? Why?

2.1 Please explain.

I want to find those rows whose values in column **num_students** are pretty low (≤ 5) because they would lead their own mean value and the standard deviation value to be meaningless when compared with other colleges. Also, If the number of students is 1, which would cause the denominator to become 0 when we calculate the standard deviation.

Actually, I want to use IQR to find outliers and remove them, but I found the result is meaningless.

2.2 Please write code in Colab to remove it and return an updated table.

2. Which row looks like an anomaly and should be removed? Why?

Find those rows whose values in column **num_students** are pretty low (≤ 5) and remove them.
Because they would lead their own mean value and the standard deviation value to be meaningless when compared with other colleges

```
[58] threshold = 5
tbl.where('num_students', are.below_or_equal_to(threshold))
```

Code Part

性別	入學管道	離校院別名稱(中)	num_students	credits_mean	credits_std	gpa_mean	gpa_std
女	學士班(指考分發/聯考)	文學院中國文學系	1	172	nan	85.0756	nan
男	繁星推薦	客家學院	1	137	nan	87.8686	nan

Output Table

Get the updated table and show it.

```
# return an updated table that have a certain feature
updated_tbl = tbl.where('num_students', are.above(threshold))

# show the entire table to check whether our result table is correct
updated_tbl.show()
```

女	學士班(指考分發/聯考)	資訊電機學院	119	150.126	14.8715	81.3947	6.56094
女	學士班申請入學	地球科學學院	134	155.507	15.7703	79.731	7.3229
女	學士班申請入學	客家學院	31	161.677	32.0639	83.1672	5.32421
女	學士班申請入學	工學院	105	152.61	15.4668	79.2026	7.36438
女	學士班申請入學	文學院	174	153.925	19.8355	80.6999	5.91329
女	學士班申請入學	理學院	88	153.739	16.8055	75.6452	7.49452

Also, we can check the # of rows of the updated table and total rows that have been removed to prove our operations is correct.

```
[63] print("# of rows after updated: ", updated_tbl.num_rows, "\n")

print("Total rows have been removed: ", tbl.num_rows - updated_tbl.num_rows)

# of rows after updated: 47

Total rows have been removed: 2
```

Note: I would use the updated table for following question.

3. Are there more girls than boys?

3.1 Please use `group` to do it.

No, there are boys more than girls. We can find out by the following table.

The # of female students is 3969 while the # of male students is 6972.

3. Are there more girls than boys?

Use `group` and `select` to find the # of boys and the # of girls.

```
# Get the table with gender column and num_students sum column
num_students_group_by_gender = updated_tbl.group('性别', sum).select('性别', 'num_students sum')
num_students_group_by_gender.show()
print()

# Get the # of boys and the # of girls to find out whether there more girls than boys.
num_female_students = num_students_group_by_gender.column('num_students sum').item(0)
num_male_students = num_students_group_by_gender.column('num_students sum').item(1)

print('Are there more girls than boys? ', num_female_students > num_male_students)
```

Code Part

性别	num_students sum
女	3969
男	6972

Output Part

Are there more girls than boys? False

3.2 What % of students are female?

36.28 % of the students are female, and 63.74% of the students are male.

Create a new column to find the percentage of students are female or male.

```
total_students = num_female_students + num_male_students

num_students_group_by_gender = num_students_group_by_gender.with_columns(
    '%', num_students_group_by_gender.column('num_students sum')/total_students
)

num_students_group_by_gender
```

Code Part

性别	num_students sum	%
女	3969	0.362764
男	6972	0.637236

Output Table

4. Which college has the most graduates?

4.1 Please use `group` to do it.

College that has the most graduates is: 管理學院 which has 2542 graduates.

4. Which college has the most graduates?

Use `group`, `select` and `sort` to find which college has the most graduates.

```
[6] # Get the table grouped by college, also with num_students sum column
num_students_group_by_college = updated_tbl.group('離校院別名稱(中)', sum).select('離校院別名稱(中)', 'num_students sum')

# Show sum of graduates of all college.
num_students_group_by_college.sort('num_students sum', descending=True).show()
```

Code Part

離校院別名稱(中)	num_students sum
管理學院	2542
工學院	2217
資訊電機學院	2045
理學院	1653
文學院	1212
地球科學學院	849
生醫理工學院	336
客家學院	87

Output Table

```
college_has_most_graduates = num_students_group_by_college.sort('num_students sum', descending=True).column(0).item(0)
num_most_graduates = num_students_group_by_college.sort('num_students sum', descending=True).column(1).item(0)

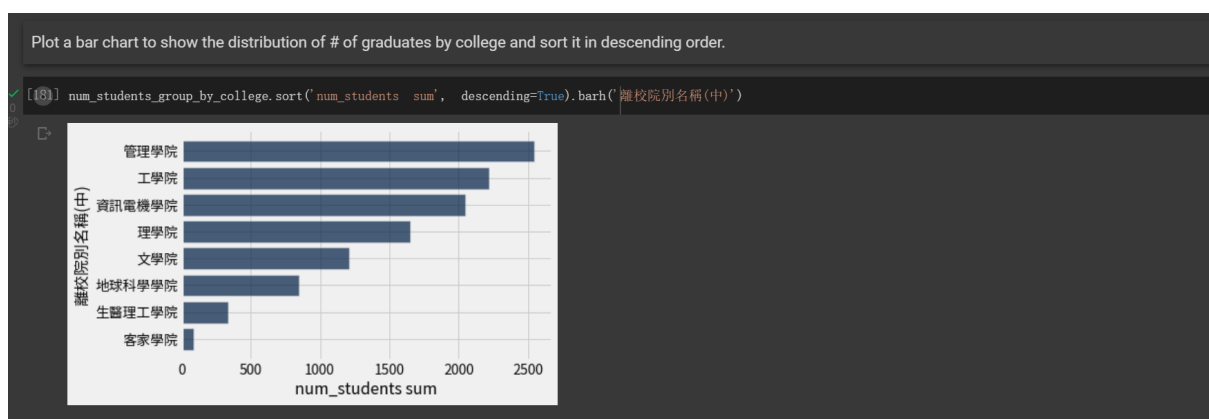
print('College that has the most graduates is:', college_has_most_graduates, ' which has ', num_most_graduates, ' graduates')
```

College that has the most graduates is: 管理學院 which has 2542 graduates

Output Answer

4.2 Please plot a bar chart to show the distribution of # of graduates by college and sort it in descending order.

I use Table method `sort` and `barh` to plot a bar chart.



5. Which college has the most unbalanced 入學管道? In other words, we say it's balanced if each 管道 has a similar number of graduates.

5.1 Please use `group` to show a pivot table, where 入學管道 and college in vertical and horizon directions, respectively, and each cell shows a total number of students.

5.2 Please plot a bar chart to show the distribution of # of graduates by college and 入學管道.

Reference

1. [Data8 docs](#)
2. [Ways to import CSV files in Google Colab - GeekforGeeks](#)