Fenglei Gu

fg1121@nyu.edu

CS 060 – Database Design and Implementation

Sep 26, 2019

Assignment 2 Summary

1. Working Environment

- Python 3.6 (64-bit)
- Encoding using UTF-8

2. Data Source

- Zhejiang Gaokao (College Entrance Exam) Admission Score Lines[1] in the First Enrollment Round of Year 2019 (浙江省 2019 年普通高校招生普通类平行投档（一段）分数线), from the official website of Zhejiang Education Examinations Authority (ZEEA)[2]. Link: https://www.zjzs.net/moban/index/402848536bab8311016bfe0466c600a8.html
- Stored locally as "浙江省 2019 年普通高校招生普通类平行投档（一段）分数线.csv" file.
- Number of records greater than 5,000.
- Dataset selected due to my personal interest (and probably many Gaokao exam candidates in Zhejiang this year).

3. Tasks

[1] The admission to college in China is (with few exceptions) based only on the scores in Gaokao (College Entrance Exam), and every year, the authority will publish the lowest Gaokao score with which one student may be admitted to any given major/division of a particular college/university.

[2] ZEEA is an authority affiliated to Department of Education of Zhejiang People's Government, who organizes Gaokao in Zhejiang Province every year, and thus is a reliable source of our data – lowest admission lines.

i. Remove comments at the bottom “注” (endnotes).

ii. Generate a new table counting how many positions are universities in each province enrolling, based on “学校代号” (University Code) (whose first two digits represent province) and “计划数” (number of enrolling).

iii. Add a column “是否双一流” (whether entitled as “Top-Ranked” by Ministry of Education of China), based on information provided in the “学校名称” (University Name) column.

iv. Modify column “学校名称” (University Name): delete their “Top-Ranked” honors tags (“一流大学建设高校” (First-Class University) or “一流学科建设高校” (First-Class Professional School)) in the name.

v. Fill the missing value (blanks) in the Column “位次” (Ranking) with 53225, which is the total number of exam candidates in this admission round. This filling is reasonable, as the blanks of ranking means the schools have not yet enrolled enough students and they still have available seats – so technically every student in this admission round who has the intention can be enrolled in. Thus, I fill these blanks as the number of total candidates in this admission round.

vi. Calculate “百分位” (Percentile) based on “位次” (Ranking) and total number of candidates.

vii. Extract the major entry with lowest admission line of each university.

viii. Remove columns “专业代号” (Major Code), “专业名称” (Major Name), “计划数” (number of students to be enrolled).

4. Results

Using the Python 3 codes I wrote as in Appendix, two charts are got in the respective .csv files. By opening them through Microsoft Excel 2016 (Figures 1 and 2), it is clear by observation that the program is working properly and giving reasonable results.

	A	B	C	D
1	Province	Number of Admission Positions		
2	Zhejiang	32802		
3	Jiangsu	2898		
4	Shanghai	2092		
5	Beijing	1838		
6	Hubei	1451		
7	Sichuan	1116		
8	Shaanxi	944		
9	Shandong	816		
10	Hunan	776		
11	Tianjin	665		
12	Guangdong	598		

Figure 1: Distribution of Zhejiang Students Admitted to Universities Located in Different Provinces

	A	B	C	D	E	F
1	学校代号	学校名称	是否双一流	分数线	位次	百分位
2	1147	清华大学	Yes	707	47	99.9117
3	1103	北京大学	Yes	707	49	99.90794
4	3131	上海交通大	Yes	699	179	99.66369
5	3102	复旦大学	Yes	697	256	99.51902
6	1104	北京大学医	Yes	694	348	99.34617
7	3171	复旦大学医	Yes	694	354	99.3349
8	3132	上海交通大	Yes	693	371	99.30296
9	1189	中国科学院	Yes	692	438	99.17708
10	3455	中国科学技	Yes	688	763	98.56646
11	1159	中国人民大	Yes	685	970	98.17755

Figure 2: Universities Ranked by Lowest Admission Line Regardless of Major

5. Conclusions0

- Completeness: I have got all the data I need – (i) how many students this year taking Gaokao in Zhejiang are enrolled to universities located in different provinces (ii) The lowest admission lines (including the score lines as well as the ranking lines) for each university, regardless of major. All universities appeared in the original datafile also appeared in the result.
- Coherency: All the data make sense as they are all within commons' expectation that: (i) most students in Zhejiang go to college in Zhejiang, Jiangsu and Shanghai; (ii) Tsinghua University and Peking University are the top 2 universities in China.
- Correctness: The values are believed to be correct as they come from the official data by the government.

- Accountability: The data can be traced back, and even if some minor modifications such as changing values are applied to the original datafile, the program will still be able to produce the respective result based on the new datafile.