Summary

Haoyu Sheng

2/13/2017

Abstract

The kane2017 package utilizes the data published in Williams Online Catalog to assess the number of students with Chinese/Cantonese/Japanese origins and compare their academic achievement with other students. This is done for 15 years from 2002 - 2016 by scanning the data directly from the text version of the webpages. A sequences of statistical comparisons will be utilized to examine the academic achievement of (partial) Asian International/American students.

Introduction

College admission, espeically recently, has been suspected of raising a higher standard towards Asian students. Though it is unclear that the admission standard has been raised at Williams, looking at the Asian students ratio of attaining Latin honors will be a useful tool in assessing the overall academic performance of Asian American students and can thus further reflect on the bar set during admission. The package kane2017 employs string manipulation(splitting, pattern detection) to split each string into its last name.

The data set is further manipulated by comparing each of the last names to a pre-existing list of possible last names of Asian origin. But processing whether if these last names are of Asian origin, we can then figure out the ratio of Latin honors among Asian students.

Data

The information used to construct the dataset is taken from the Registrar office. The text file is manually converted from the catalog through coying and pasting.Then the text file is converted into Data Frames with a column indicating the year after each student's name. In the end, I compile a data frame "DataSummary.RData" that includes all entries of names from 2002-2016 and the years associated with all the names.

In addition, the Chinese last name list is compiled through all possible pinyin combinations provided by linked phrase, the Cantonese last name list is provided by linked phrase and the Japanese last names are provided by linked phrase.

Though the input are quite concise, since last names and honors are the only things we need, we need to cut down the strings.

For example, for input: *+Kathleen Malone Palmer, with highest honors in Neuroscience, readHonor will use

unlist(str_split(data[j],","))

gsub(".*", "", name)

```
  to turn the original string into Palmer.
```

After obtaining a list of last names, we then compare the last names with the given list of last names. Now we have chineseLastName, japaneseLastName, cantoneseLastName, and totalVeri, which is the collection of the previous three.

The verifiers and student entries are previously stored through RData and passed into readHonor(data, verifier) readHonor will then return a data.frame as below

```
readHonor(DataSummary, chineseLastName)[1:5,]
```

```
##   name honor year sigmaXi phiBetaKappa
## 1 Fang Magna 2003   FALSE        FALSE
## 2  Tan Magna 2003    TRUE        FALSE
## 3 Wang Magna 2003    TRUE         TRUE
## 4   Yi Magna 2003    TRUE         TRUE
## 5 Chen       2003   FALSE        FALSE
```

Use readHonor

The information stored within the locally stored RData files of student names and academic honors can be accessed through the **readHonor** method. This function outputs a dataframe containing the name, honor, year, sigmaXi and phiBetaKappa result for all students from 2002 - 2016.

For easier user interaction, I consolidated the data files into **DataSummary.RData**, which includes all information through out all 15 years. The input files are stored based on individual years as 2002-2016.RData if users want to experience the function by using single year data.

<- data: input data set that contains the list of student names as provided by the registrar office.

<- verifier: input data set that contains the possible last names of a givevn cultural origin. This package includes cantonese, chinese, and japanese

Use statsummary

The function statsummary is used to generate different analytical graphical representation of the honor data collected by **readHonor(data, verifier)**. The function takes in one parameter that indicates the type of analysis desired.

> statsummary(type)

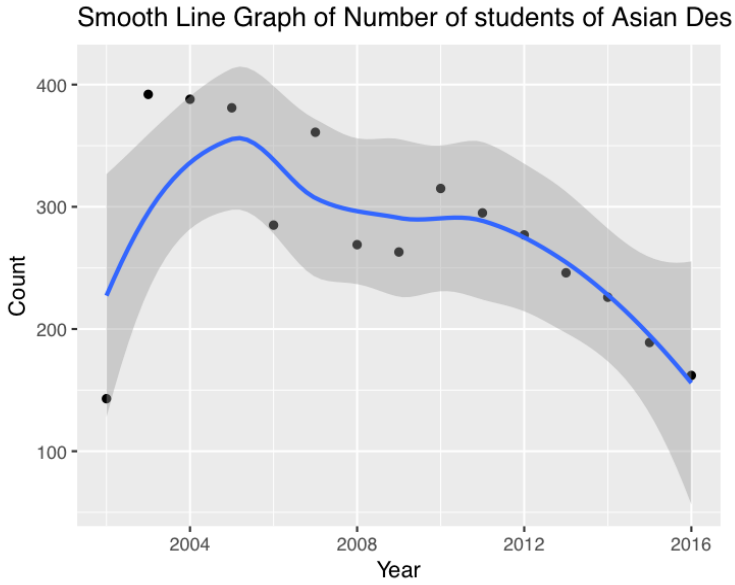
The parameter type has 5 different input values: breakdown, histogram, summary by origin, summary by percentage, summary by account

```
statsummary("breakdown")
```

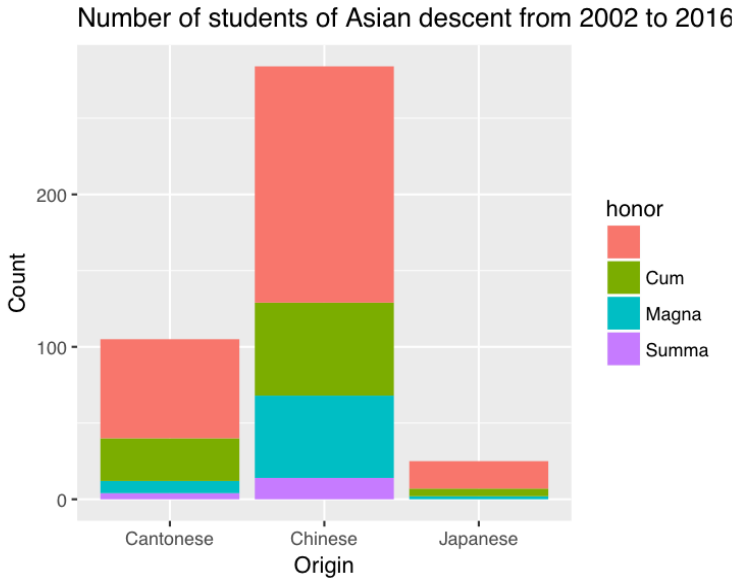
```
##    year total summa magna cum sigmaXi phiBetaKappa
## 1  2002   143     1     3   3       4            4
## 2  2003   392     0     6   1       5            3
## 3  2004   388     1     1   4       2            1
## 4  2005   381     1     3   3       3            2
## 5  2006   285     0     0   4       3            0
## 6  2007   361     1     3   7       6            1
## 7  2008   269     1     2   6       5            3
## 8  2009   263     1     6   8      10            6
## 9  2010   315     1     4   9       9            9
## 10 2011   295     1     7   5      10            8
## 11 2012   277     3     7   8       9            8
## 12 2013   246     2     4  12       6            4
## 13 2014   226     4     9   9      13           10
## 14 2015   189     1     7   8      10            6
## 15 2016   162     0     2   7       5            2
```

```
statsummary("histogram")
```

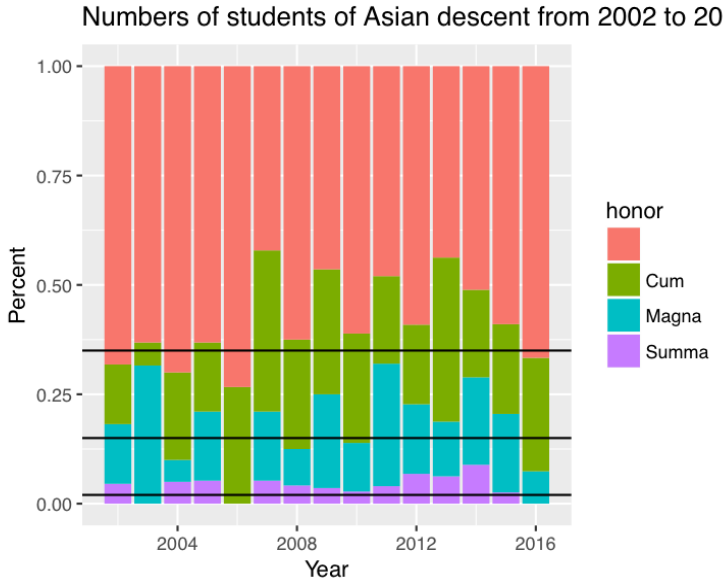
```
## `geom_smooth()` using method = 'loess'
```



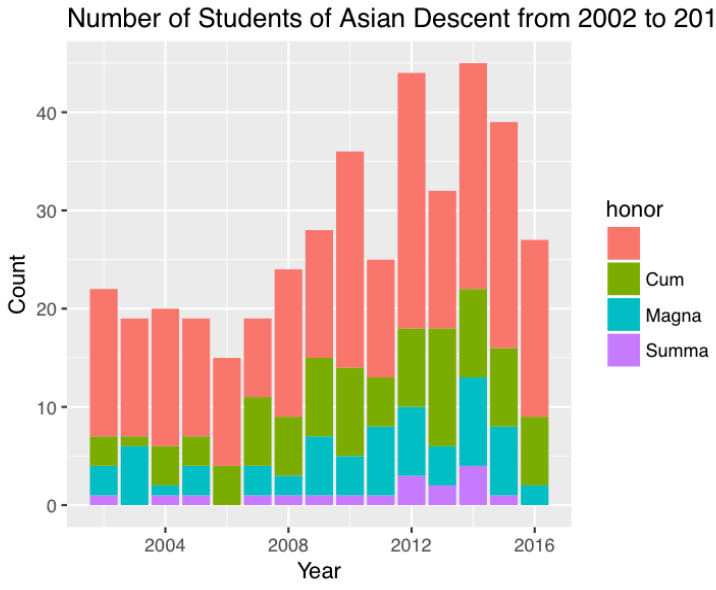

```
statsummary("summary by origin")
```



```
statsummary("summary by percentage")
```




```
statsummary("summary by count")
```



‘‘

Conclusion

The summary statstics indicates that according to the Asian last name reference sheets, students of Asian origins tend to be significantly overachieving, with an overall average Cum Laude percentage around 7.5% higher than the average 35%. Also, in the scatterplot graphs, the majority of all stats over years is higher than the average ratio of that particualr honor and above. Thus, it can be inferred that Asian students tend to be more academically achieving. Such conclusion, therefore, should serve as a powerful evidence towards some of the variations in admission standards Asian students might be facing.

However, this package is still incomplete in many ways. First, distinguishing last names from its linguistic composition can be extremely ambiguous. It doesn't take into account some of the identical last names different cultures can share. Also, it doesn't take into account students of mixed racial identity. Therefore, the next step will be to incorporate more last name recognizing templates as well as trying to distinguish last names of ambiguous identities base on first names.