141A HW2

Man Pan

Student ID: 914656278

Email: manpan@ucdavis.edu

1. **Are there any features with no missing values? Which features have the most missing values? Explore the missing values and report any patterns you find.**

Features with no missing values:

Number_of_missing_value	Features
0	id
0	ope8_id
0	ope6_id
0	name
0	city
0	state
0	degrees_awarded.predominant
0	degrees_awarded.highest
0	ownership
0	main_campus
0	branches
0	institutional_characteristics.level
0	zip
0	academic_year

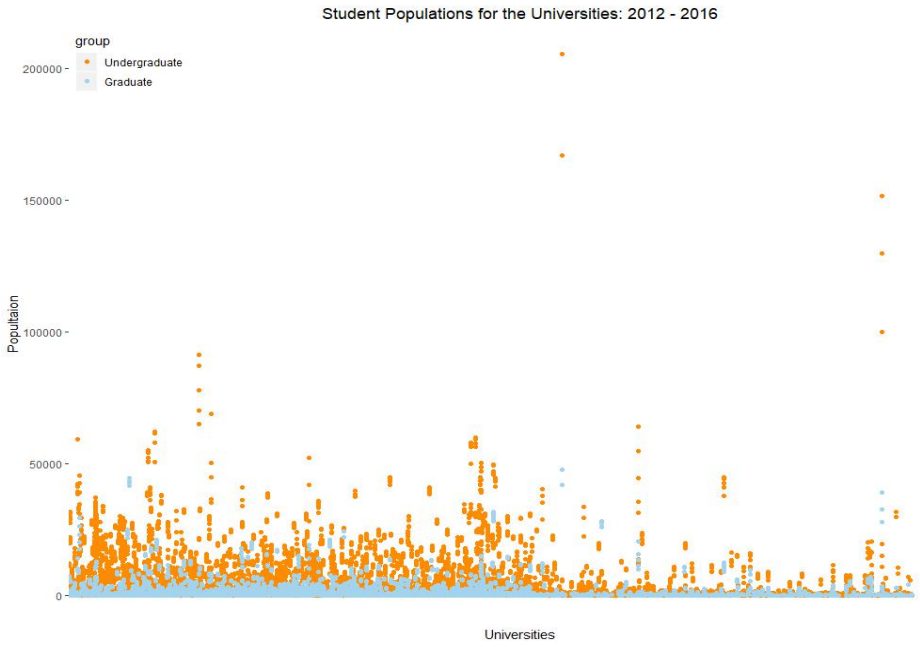
features have the most missing values:

Number_of_missing_value	Features
38068	minority_serving.historically_black
38068	minority_serving.predominantly_black
38068	minority_serving.annh
38068	minority_serving.tribal
38068	minority_serving.aanipi
38068	minority_serving.hispanic
38068	minority_serving.nant
38068	men_only
38068	women_only
38068	operating

I find the features with no missing data are the basic information like id,name,zip..... Those information can get easily, so we do not have missing value. However, 38068 missing values mean that we have no information about those features. Those features are all details that are harder to get. Those features are useless in the following research.

2. Explore student populations for the universities. Are there any schools with unusual populations? What is the relationship between undergraduate and graduate populations? Are there exceptions to the relationship?

From the document, we can know feature “size” is the undergraduate population, and feature “grad_students” is the graduate population. We can plot the populations grouped by graduate and undergraduate for universities to explore some useful information. (In this plot, I do not split the population by different academic year)

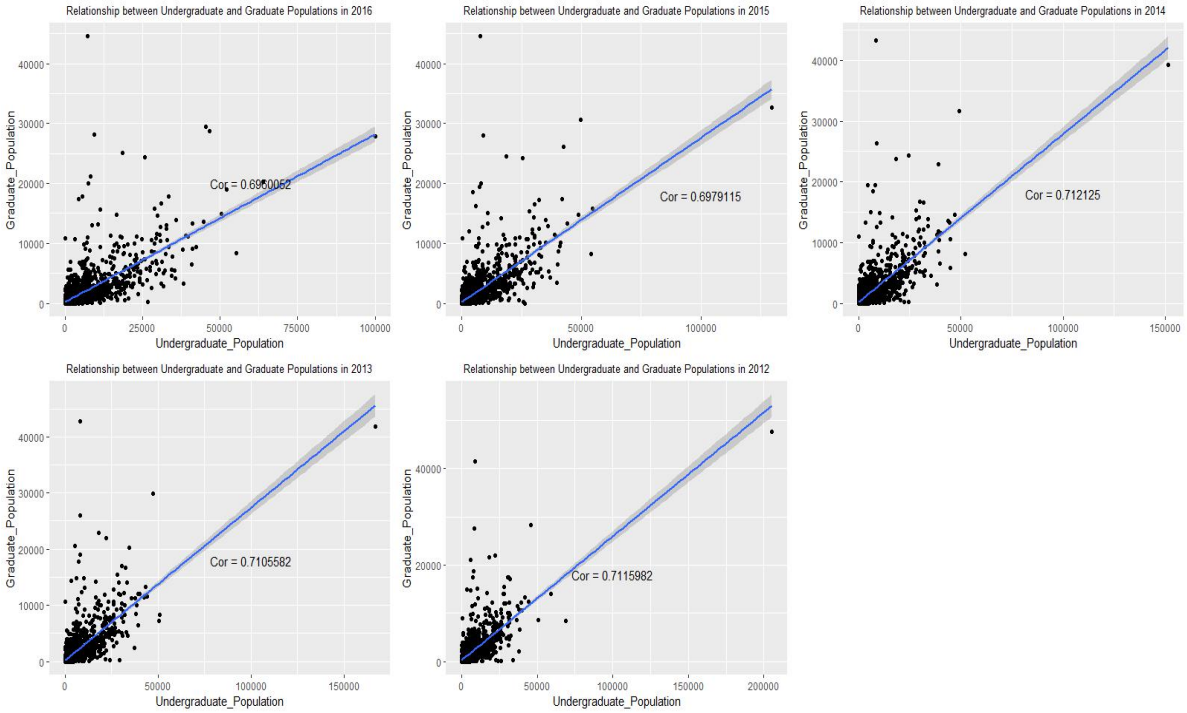


There are 3 universities with unusual populations for undergraduate population:
University of Phoenix-Online Campus, University of Phoenix-Arizona,
Ivy Tech Community College
there are 4 universities with unusual populations for graduate population:
Walden University, University of Phoenix-Online Campus, University of Phoenix-Arizona,
Capella University
Then, I add undergraduate and graduate population to one column as total population. Plot total population grouped by different academic year as follow:

From the plot, we can conclude:
In 2012, University of Phoenix-Online Campus has unusual population. The total population is 252946;

In 2013, University of Phoenix-Online Campus has unusual population. The total population is 208716;
In 2014, University of Phoenix-Arizona has unusual population. The total population is 190745;
In 2015, University of Phoenix-Arizona has unusual population. The total population is 162361;
In 2016, University of Phoenix-Arizona has unusual population. The total population is 127929;
They are all have a larger unusual population.

Finally, I select the population data in 2012-2016 to explore the relation between Undergraduate and graduate populations.



From the plot, we can find that the correlation is 0.6960052 (2016), 0.6979115 (2015), 0.712125 (2014), 0.7105582 (2013), 0.7115982 (2012), which means undergraduate and graduate population are positively correlated. I think there is no exception to the relationship. They just have different coefficient, which means they have different levels of positive correlation.

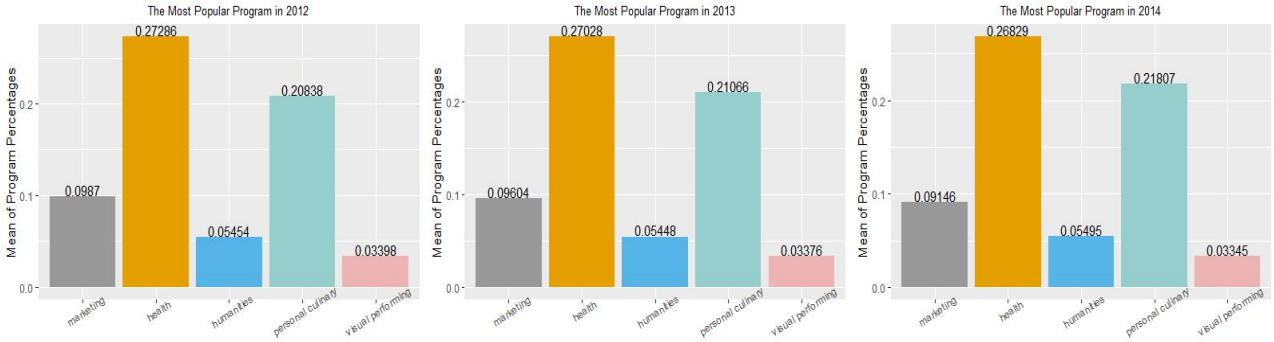
3. **Explore the program percentages for the universities. What programs are the most popular? What programs are the least popular? Are there any program percentages that show patterns different from the others?**

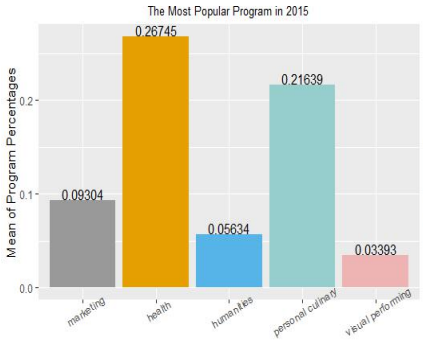
I use the mean of the program percentage to make sure which programs are the most popular and the least popular. Use the 2012 year data as an example:

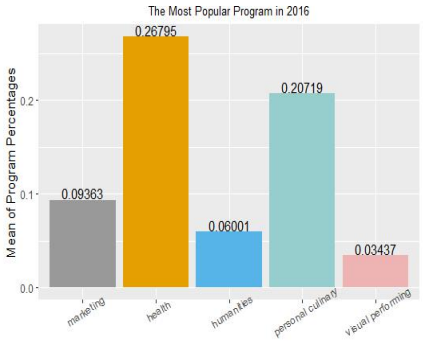
```
Programe[order(Programe$Mean_of_program_percentages),]    # sorting by increasing
 Mean_of_program_percentages                                          Features
                5.174410e-05                        program_percentage.library
                6.167208e-05                       program_percentage.military
                5.934473e-04             program_percentage.science_technology
                1.218599e-03          program_percentage.ethnic_cultural_gender
                1.469425e-03                       program_percentage.architecture
```

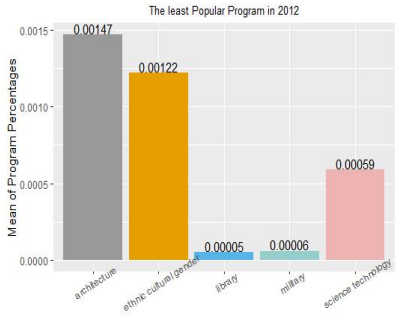
```
Programe[order(-Programe$Mean_of_program_percentages),]  # sorting by decreasing
 Mean_of_program_percentages                                  Features
                   2.728638e-01                 program_percentage.health
                   2.083759e-01       program_percentage.personal_culinary
                   9.869936e-02    program_percentage.business_marketing
                   5.454186e-02              program_percentage.humanities
                   3.398180e-02      program_percentage.visual_performing
```

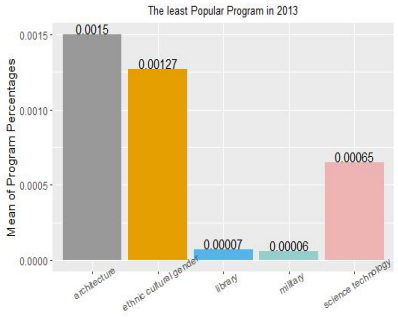
Then, I plot 5 graphs in each year to present the mean of program percentages (just select the top 5 most/least popular programs).

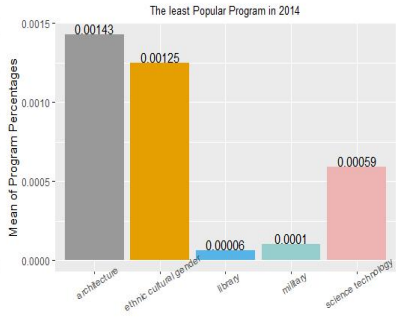


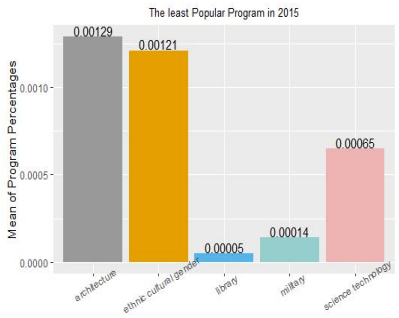


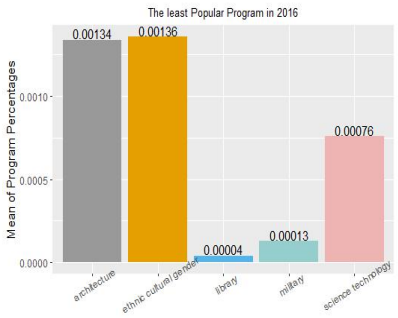








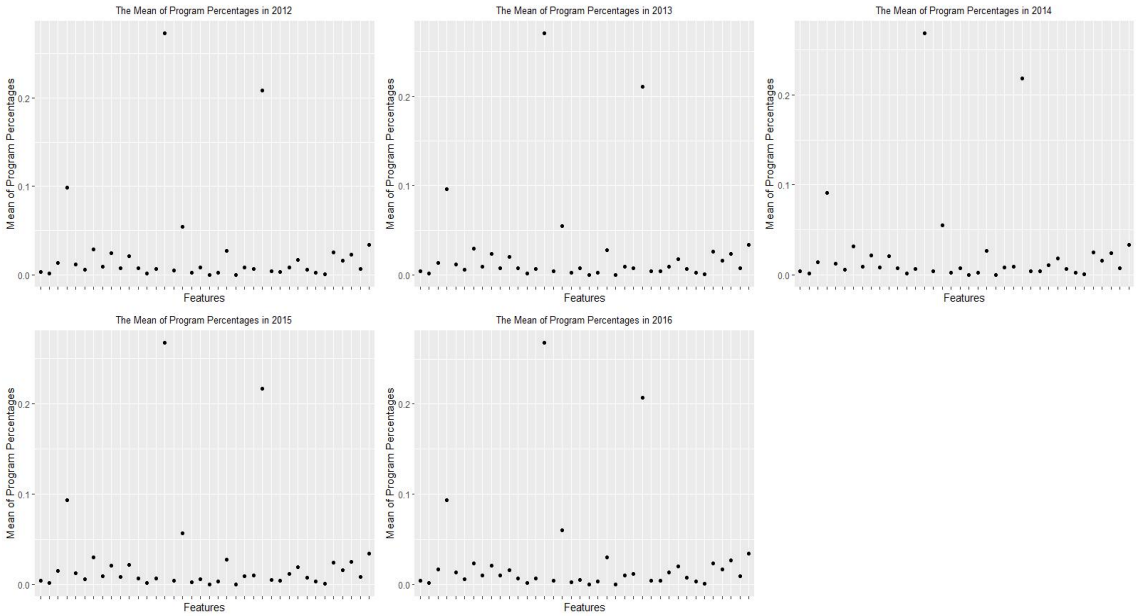




Conclusion:

the most popular programs are : Health, personal_culinary , business_marketing , humanities,

visual_performing. The least popular programs are: Library, military, science_technology, ethnic_cultural_gender, architecture.



From the scatter plot above, I can find some program percentages that show patterns different from the others. They are Health, personal_culinary , business_marketing , humanities, which are all have large program percentage compared to others.

4. **How does tuition vary across different states? Is there a relationship between the number of universities in a state and tuition? Do these characteristics differ for in-state tuition and out-of-state tuition?**

I add the "tuition.in_state" and "tuition.out_of_state" into one column named “total_tuition”. I calculate the variance of total tuition according to different states and different academic year.

We can find the state which has the most variance of total tuition has the most diverse. The opposite is also true.

The 5 most diverse state (total tuition): 2012 -2016

The 5 least diverse state(total tuition): 2012 - 2016

	state	Var_tuition	academic_year
1	CT	704647936	2012
2	DC	870868533	2012
3	ME	725367601	2012
4	MA	795959278	2012
5	VT	678110125	2012
6	CT	704647936	2013
7	DC	870868533	2013
8	ME	725367601	2013
9	MA	795959278	2013
10	VT	678110125	2013
11	CT	704647936	2014
12	DC	870868533	2014
13	ME	725367601	2014
14	MA	795959278	2014
15	VT	678110125	2014
16	CT	704647936	2015
17	DC	870868533	2015
18	ME	725367601	2015
19	MA	795959278	2015
20	VT	678110125	2015
21	CT	704647936	2016
22	DC	870868533	2016
23	ME	725367601	2016
24	MA	795959278	2016
25	VT	678110125	2016

	state	Var_tuition	academic_year
1	AS	23000	2012
2	MP	536609	2012
3	PW	0	2012
4	VI	854053	2012
5	MH	41720	2012
6	AS	23000	2013
7	MP	536609	2013
8	PW	0	2013
9	VI	854053	2013
10	MH	41720	2013
11	AS	23000	2014
12	MP	536609	2014
13	PW	0	2014
14	VI	854053	2014
15	MH	41720	2014
16	AS	23000	2015
17	MP	536609	2015
18	PW	0	2015
19	VI	854053	2015
20	MH	41720	2015
21	AS	23000	2016
22	MP	536609	2016
23	PW	0	2016
24	VI	854053	2016
25	MH	41720	2016

We can find DC, MA, ME, CT, VT states have the most diverse in total tuition;
AS, MP, PW, VI, MH states have the least diverse in total tuition.

I use the same method to find the most/least diverse state according to in-state tuition and out-of-state tuition.

The 5 most diverse state (in state tuition):
2012 -2016

The 5 least diverse state(in state tuition):
2012 - 2016

	state	Var_instate	academic_year
1	CT	207299635	2012
2	DC	211246306	2012
3	MA	219157120	2012
4	RI	212956026	2012
5	VT	207330629	2012
6	CT	207299635	2013
7	DC	211246306	2013
8	MA	219157120	2013
9	RI	212956026	2013
10	VT	207330629	2013
11	CT	207299635	2014
12	DC	211246306	2014
13	MA	219157120	2014
14	RI	212956026	2014
15	VT	207330629	2014
16	CT	207299635	2015
17	DC	211246306	2015
18	MA	219157120	2015
19	RI	212956026	2015
20	VT	207330629	2015
21	CT	207299635	2016
22	DC	211246306	2016
23	MA	219157120	2016
24	RI	212956026	2016
25	VT	207330629	2016

	state	Var_instate	academic_year
1	AS	4500	2012
2	MP	296705	2012
3	PW	0	2012
4	VI	53417	2012
5	MH	4600	2012
6	AS	4500	2013
7	MP	296705	2013
8	PW	0	2013
9	VI	53417	2013
10	MH	4600	2013
11	AS	4500	2014
12	MP	296705	2014
13	PW	0	2014
14	VI	53417	2014
15	MH	4600	2014
16	AS	4500	2015
17	MP	296705	2015
18	PW	0	2015
19	VI	53417	2015
20	MH	4600	2015
21	AS	4500	2016
22	MP	296705	2016
23	PW	0	2016
24	VI	53417	2016
25	MH	4600	2016

For in state tuition, we find CT, DC, MA, RI, VT states have the most diverse;
AS, MP, PW, VI ,MH states have the least diverse.

The 5 most diverse state (out of state tuition): 2012 -2016

The 5 least diverse state(out of state tuition): 2012 - 2016

	state	Var_outstate	academic_year
1	CT	158309566	2012
2	DC	210557876	2012
3	ME	168466229	2012
4	MD	154014078	2012
5	MA	172546184	2012
6	CT	158309566	2013
7	DC	210557876	2013
8	ME	168466229	2013
9	MD	154014078	2013
10	MA	172546184	2013
11	CT	158309566	2014
12	DC	210557876	2014
13	ME	168466229	2014
14	MD	154014078	2014
15	MA	172546184	2014
16	CT	158309566	2015
17	DC	210557876	2015
18	ME	168466229	2015
19	MD	154014078	2015
20	MA	172546184	2015
21	CT	158309566	2016
22	DC	210557876	2016
23	ME	168466229	2016
24	MD	154014078	2016
25	MA	172546184	2016

	state	Var_outstate	academic_year
1	AS	13250	2012
2	MP	35280	2012
3	FM	377817	2012
4	PW	0	2012
5	MH	23520	2012
6	AS	13250	2013
7	MP	35280	2013
8	FM	377817	2013
9	PW	0	2013
10	MH	23520	2013
11	AS	13250	2014
12	MP	35280	2014
13	FM	377817	2014
14	PW	0	2014
15	MH	23520	2014
16	AS	13250	2015
17	MP	35280	2015
18	FM	377817	2015
19	PW	0	2015
20	MH	23520	2015
21	AS	13250	2016
22	MP	35280	2016
23	FM	377817	2016
24	PW	0	2016
25	MH	23520	2016

For out of state tuition, we find CT, DC, MA, ME, MD states have the most diverse;
AS, MP, PW, FM ,MH states have the least diverse, which have slightly difference compared to those of total tuition.

Relationship between the number of universities in a state and tuition

(From the answer above, I find the academic year does not effect diverse of the tuition. So I do not take academic year into account below.)

According to relationship between the number of universities in a state and total tuition, I find PW, AS, MP, FM, MH states have the least total tuition, and they are all just have 5 universities. However, NH,DC,MA,VT,RI states have the most total tuition. Their number of universities in state are 206, 124, 984,140, 130 separately, which are not the largest number of universities in state. Thus, we can conclude that there is some relationship between the number of universities in a state and the lower tuition. The similar characteristics for in-state tuition and out-of-state tuition.

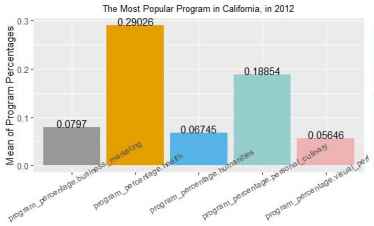
5. **Which colleges have the most diverse demographics? Make sure to explain how you measured ‘‘diversity’’ for this problem, in addition to discussing your conclusions .**

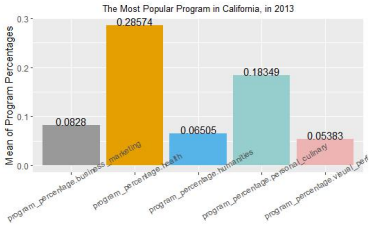
Firstly, I choose "demographics.men" and "demographics.women", then, using demographics.men /demographics.women (ratio) to as the first criterion to measure “diversity”. If the ratios are close to or equal to 1, the schools would have the most diverse demographics. I choose some schools that have the most diverse demographics as follow:

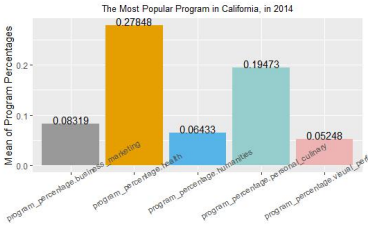
name	demographics.men	demographics.women	ratio
Mid-Atlantic Christian University	0.5000	0.5000	1.00000
Moler Barber College	0.5000	0.5000	1.00000
Jones Hair Design College	0.5000	0.5000	1.00000
Christian Life College	0.5000	0.5000	1.00000
Northeast Technology Center-Pryor	0.5000	0.5000	1.00000
International Baptist College and Seminary	0.5000	0.5000	1.00000
Hood Theological Seminary	0.5000	0.5000	1.00000
Charlie's Guard-Detective Bureau and Academy Inc	0.5000	0.5000	1.00000

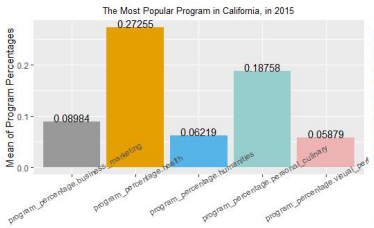
Then, I select some race ethnicity variables : white, black, Hispanic, Asian, Aian, Nhpi, unknown to calculate the variance of each school as another criterion to find the college which has the most variance of the percentage of race. From the output, Marsha Kay Beauty College, New Tyler Barber College Inc , Shorter College, Velvatex College of Beauty Culture, CET-El Centro..... have the most variance of the percentage of race 0.14286. (I just select 5 as examples).

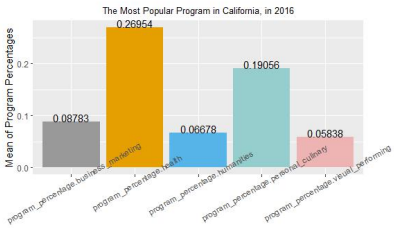
6.(a)Compare program percentage in California colleges.(Find the most/least popular program in California)

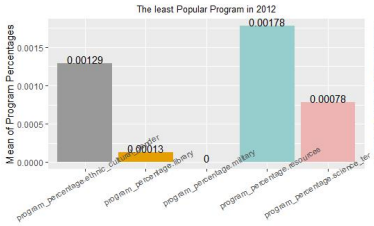


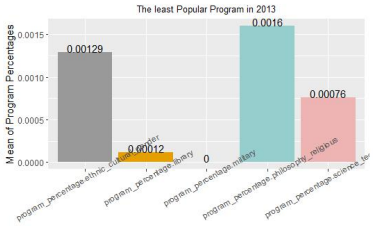


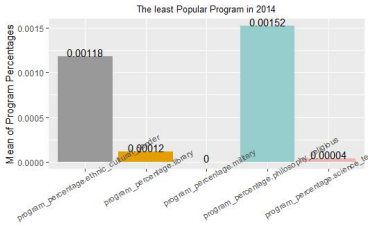


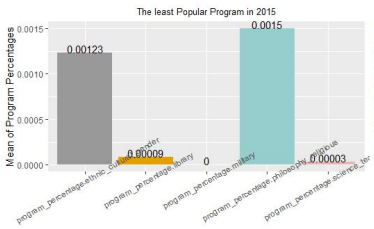


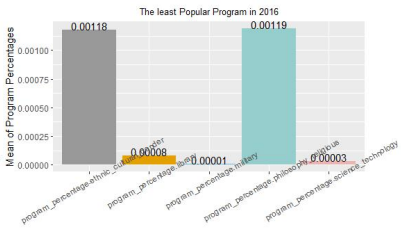






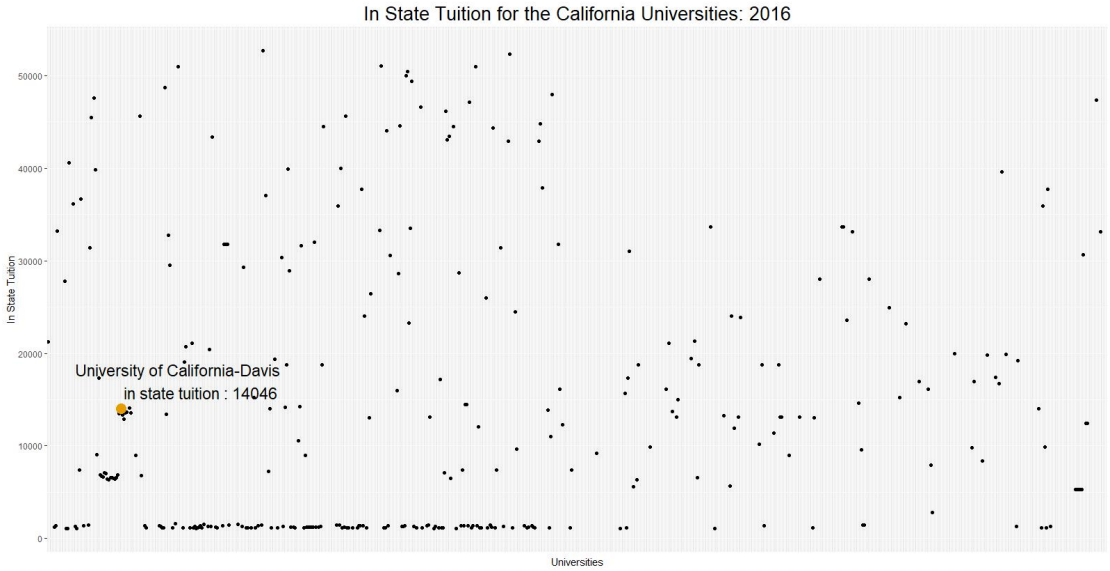


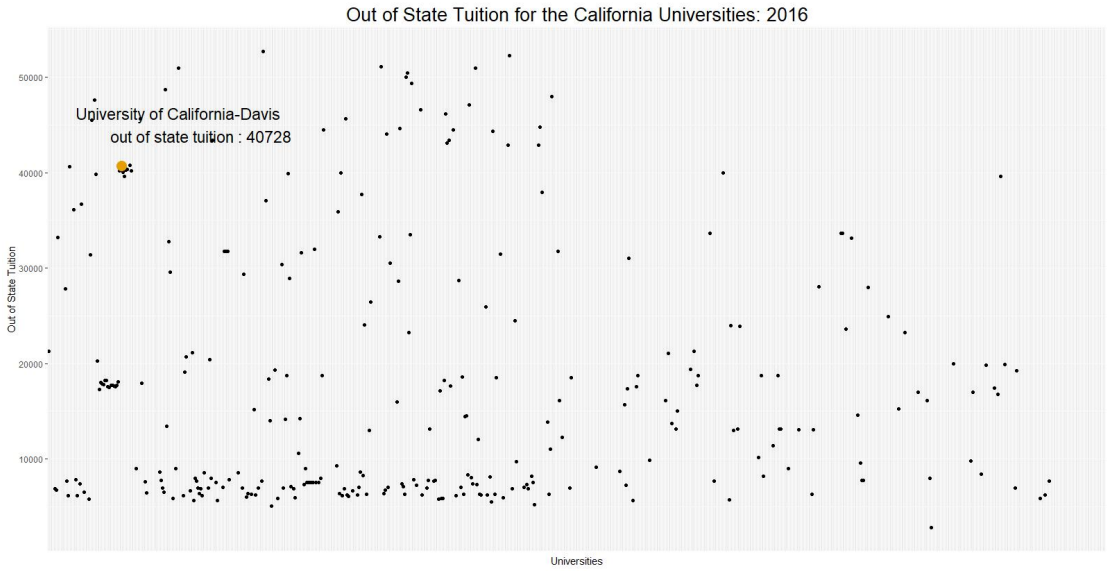




Conclusion：In California, the most popular programs are humanities, visual_performing, health, business_marketing , personal_culinary; the least popular programs are resources, ethnic_cultural_gender, library, military, science_technology, philosophy_religious, science_technology.

(b) Compare tuition of UCD with that of other schools in California states in 2016.





Conclusion: From the outputs, I find: for in state tuition, UCD's tuition is lower than the mean level(14625). However, for out of state tuition, UCD's tuition is much higher than the mean level(18376).

7. Reflect on the questions you answered in Problem 6. Did they lead to interesting conclusions? Why or why not? Did they raise new questions? Is it the question that makes a result interesting, the data, or both? Explain.

For the most popular program, the most popular programs in California are humanities, visual_performing, health, business_marketing , personal_culinary, while the most popular programs in US are Health, personal_culinary , business_marketing , humanities, visual_performing. They are the same. Years do not effect the most popular program.

For the least popular program, the least popular programs in California are resources, et hnic_cultural_gender, library, military, science_technology, philosophy_religious, science_tec hnology, while the least popular programs in US are Library, military, science_technology, ethnic_cultural_gender, architecture. They are different. And different year has different l east popular programs.

new question: Are the most popular programs the same in other State? Are there some schools that do not have these popular programs?

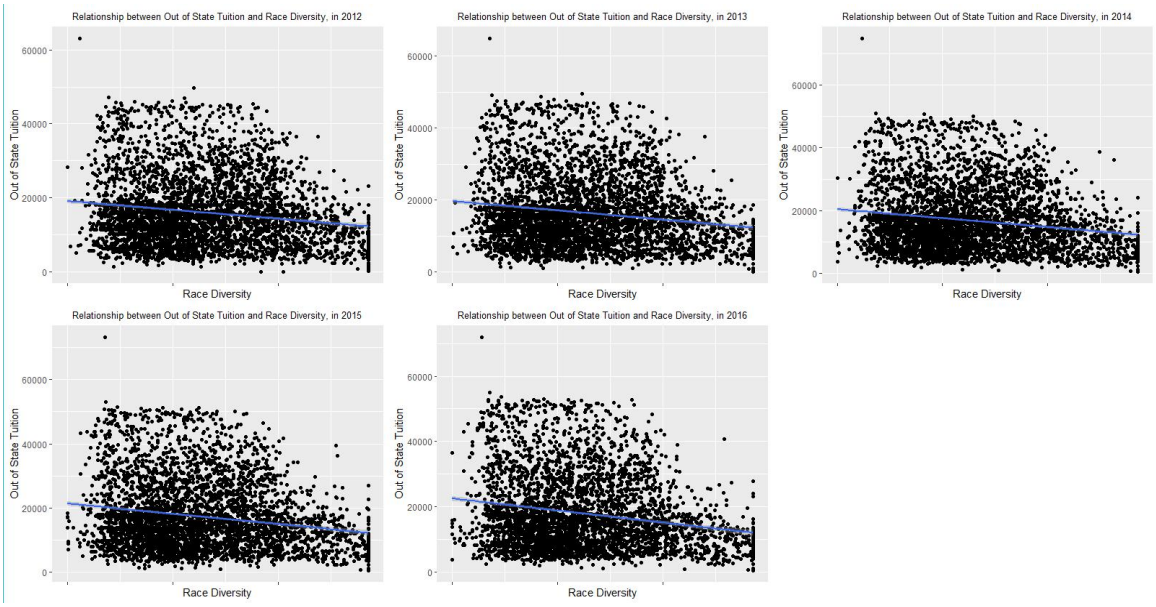
From the tuition outputs, I find out of state tuition in UCD are much higher compared to other schools in California. It will be interesting to explore the relationship between out of state tuition and the race diversity in California universities. Maybe, California already has so many non-resident students, UCD does not need to decrease the tuition to attract more out of state students.

New question: Is there a relationship between the diversity of race and out of state tuition in California?

8. List and answer 2 follow-up questions raised by any of the work you did for this assignment. Along with each question and answer, make sure to explain what raised the question for you.

(a) Is there a relationship between the diversity of race and out of state tuition ?

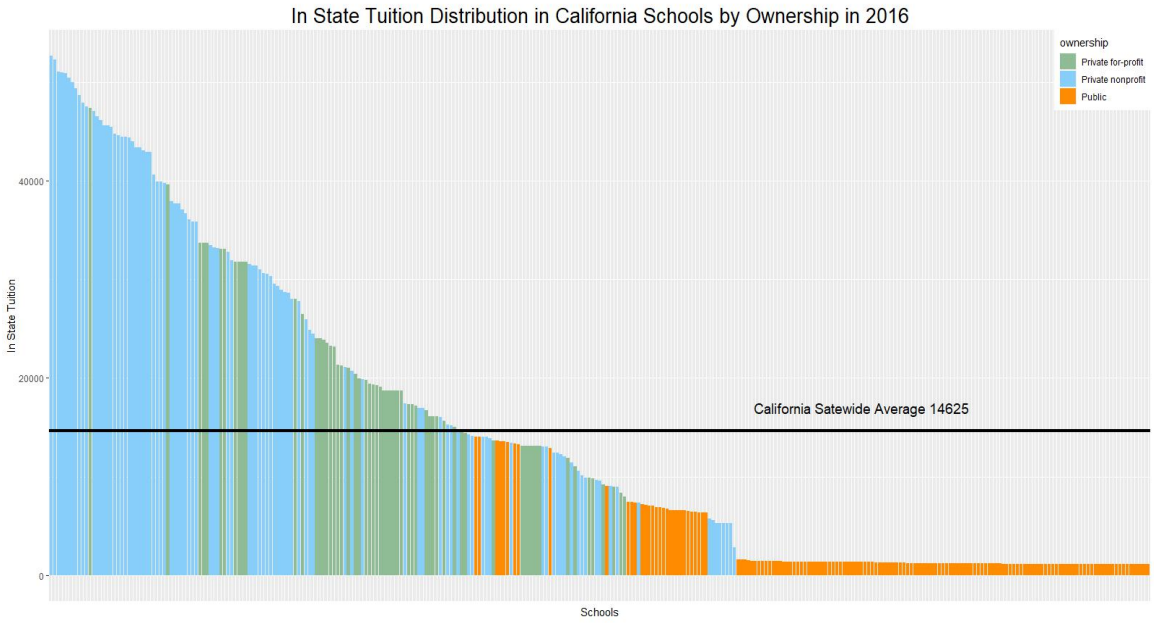
I find UCD out of state tuition is much higher than other universities', while UCD in state tuition is lower. And I know, there are so many national students in UCD. So I want to explore the relationship between race diversity and out of state tuition.



From the output, I find there are just slightly negative correlation between the race diversity and out of state tuition in each year. Form 2012 to 2016, the correlation of race diversity and out of state tuition are -0.16720, -0.17135,-0.18209,-0.19415,-0.20903 separately.

(b) the variation in in-state-tuition in California schools by ownership in 2016?

When I explore the relationship between the number of universities in a state and tuition , I did not find a highly relation. Then, I think the in state tuition maybe related to the ownership.



The graph above shows the variation in in_state_tuition in California 2016 and by ownership. Private nonprofit schools' tuition are much higher than others. Public schools are the lowest in state tuition among the three ownership. And we know the California statewide average in state tuition is 14625. No public schools' tuition is higher than the average level. Also, private for profit schools tend to have in state tuition around the California statewide average of 14625.

Appendix

```
data = readRDS("college_scorecard.rds")
####### 1
# calculate missing value of each column and add the features' name
res = NULL
for (i in 1:ncol(data)){
      temp<-sum(is.na (data[,i]))
      temp<-as.data.frame(temp)
      temp$Features<-colnames(data)[i]
      res<-rbind(res,temp)
  }
names(res)[names(res) == "temp"] = "Number_of_missing_value"
res[order(res$Number_of_missing_value),]      # sorting by increasing
res[order(-res$Number_of_missing_value),]    # sorting by decreasing

######### 2
options("scipen"=100, "digits"=5)
# undergraduate population
a = data[,c("id","size")]
a$size = as.numeric(a$size)
# graduate population
b = data[,c("id","grad_students")]
b$grad_students = as.numeric(b$grad_students)

# combine the two population in one dataset
group = rep(0, nrow(a))    # group equals to 0 means undergraduate
a = cbind(a, group)
group = rep(1, nrow(b))    #    group equals to 1 means graduate
b = cbind(b, group)
names(a)[names(a) == "size"] = "Population"
names(b)[names(b) == "grad_students"] = "Population"
ab = rbind(a,b)
ab$group = as.factor(ab$group)
ab$id = as.factor(ab$id)

levels(ab$group)[which(levels(ab$group)==0)] = "Undergraduate"
levels(ab$group)[which(levels(ab$group)==1)] = "Graduate"
ab$group = as.factor(ab$group)

library(ggplot2)
# Basic scatter plot of undergraduate and graduate population
ggplot(ab, aes(x = id, y = Population, color = group)) + geom_point()+
  scale_color_manual(values = c("Graduate"="lightskyblue2", "Undergraduate"="darkorange"))+
  theme(axis.ticks.x = element_blank() )+theme(plot.title=element_text(hjust=0.5))+
```

```
  labs(title="Student Populations for the Universities: 2012 - 2016",x="Universities",y="Popultaion")+
  theme(axis.text.x = element_blank() ) +
  theme(legend.position=c(0,1), legend.justification=c(0,1))

# undergraduate unusual populations
data[data[,"id"] == '372213',]$name    #University of Phoenix-Online Campus
ab[ab[,"id"] == '372213',]$Population #    205286 166816    47660    41900
data[data[,"id"] == '484613',]$name    #University of Phoenix-Arizona
ab[ab[,"id"] == '484613',]$Population    # 151558 129615 100011    39187    32746    27918
data[data[,"id"] == '150987',]$name    #Ivy Tech Community College
ab[ab[,"id"] == '150987',]$Population # 91112 87017 77657 70074 65092

# graduate unusual populations
data[data[,"id"] == '125231',]$name #Walden University
ab[ab[,"id"] == '125231',]$Population #8558    8064    8658    7815    7329 41513 42811 43228 44560 44530
data[data[,"id"] == '372213',]$name #University of Phoenix-Online Campus
ab[ab[,"id"] == '372213',]$Population #205286 166816    47660    41900
data[data[,"id"] == '484613',]$name #University of Phoenix-Arizona
ab[ab[,"id"] == '484613',]$Population #151558 129615 100011    39187    32746    27918
data[data[,"id"] == '413413',]$name #Capella University
ab[ab[,"id"] == '413413',]$Population #8066    7935    8738    9098    9385 27667 26060 26311 27969 28176

##### population grouped by different academic year
c = data[,c("id","size","grad_students","academic_year")]
c$size[is.na(c$size)] = 0
c$grad_students[is.na(c$grad_students)] = 0

c$total_population = c$size + c$grad_students
c$total_population[c$total_population == 0] = NA
c$id = as.factor(c$id)
c$academic_year = as.factor(c$academic_year)
levels(c$academic_year)

hist6 = ggplot(c, aes(x = id, y = total_population, color = academic_year)) + geom_point()+
  scale_color_manual(values = c("2012"="springgreen3", "2013"="darkorange", "2014"="lightcoral", "2015"="violetred2", "2016"="mediumpurple3"))+
  theme(axis.ticks.x = element_blank() )+ theme(plot.title=element_text(hjust=0.5))+
  labs(title="Student Populations for the Universities of Different Years",x="Universities",y="Total Popultaion")+
  theme(axis.text.x = element_blank() ) +theme(plot.title = element_text(size = 10))+
  theme(legend.position=c(1,1), legend.justification=c(1,1))
```

```
c_2012 = c[c[,'academic_year'] == "2012",]
c_2013 = c[c[,'academic_year'] == "2013",]
c_2014 = c[c[,'academic_year'] == "2014",]
c_2015 = c[c[,'academic_year'] == "2015",]
c_2016 = c[c[,'academic_year'] == "2016",]
hist1 = ggplot(c_2012, aes(x = id, y = total_population, color = academic_year )) + geom_point()+
  scale_color_manual(values = c("2012"="palevioletred1"))+
  theme(axis.ticks.x = element_blank() )+ theme(plot.title=element_text(hjust=0.5))+
  labs(title="Student Populations for the Universities: 2012",x="Universities",y="Total Popultaion")+
  theme(axis.text.x = element_blank() )+theme(plot.title = element_text(size = 10))+
  theme(legend.position = "none")
hist2 = ggplot(c_2013, aes(x = id, y = total_population, color = academic_year )) + geom_point()+
  scale_color_manual(values = c("2013"="violet"))+
  theme(axis.ticks.x = element_blank() )+ theme(plot.title=element_text(hjust=0.5))+
  labs(title="Student Populations for the Universities: 2013",x="Universities",y="Total Popultaion")+
  theme(axis.text.x = element_blank() ) +theme(plot.title = element_text(size = 10))+
  theme(legend.position = "none")
hist3 = ggplot(c_2014, aes(x = id, y = total_population, color = academic_year )) + geom_point()+
  scale_color_manual(values = c("2014"="steelblue1"))+
  theme(axis.ticks.x = element_blank() )+ theme(plot.title=element_text(hjust=0.5))+
  labs(title="Student Populations for the Universities: 2014",x="Universities",y="Total Popultaion")+
  theme(axis.text.x = element_blank() ) +theme(plot.title = element_text(size = 10))+
  theme(legend.position = "none")
hist4 = ggplot(c_2015, aes(x = id, y = total_population, color = academic_year )) + geom_point()+
  scale_color_manual(values = c("2015"="tan1"))+
  theme(axis.ticks.x = element_blank() )+ theme(plot.title=element_text(hjust=0.5))+
  labs(title="Student Populations for the Universities: 2015",x="Universities",y="Total Popultaion")+
  theme(axis.text.x = element_blank() ) +theme(plot.title = element_text(size = 10))+
  theme(legend.position = "none")
hist5 = ggplot(c_2016, aes(x = id, y = total_population, color = academic_year )) + geom_point()+
  scale_color_manual(values = c("2016"="paleturquoise3"))+
  theme(axis.ticks.x = element_blank() )+ theme(plot.title=element_text(hjust=0.5))+
  labs(title="Student Populations for the Universities: 2016",x="Universities",y="Total Popultaion")+
  theme(axis.text.x = element_blank() ) +theme(plot.title = element_text(size = 10))+
  theme(legend.position = "none")

library(gridExtra)
grid.arrange(hist1,hist2,hist3,hist4,hist5,hist6,nrow=2)
```

```
# get the unusual population each year
c_2012[c_2012[,"id"] == '372213',]$total_population
c_2013[c_2013[,"id"] == '372213',]$total_population
c_2014[c_2014[,"id"] == '484613',]$total_population
c_2015[c_2015[,"id"] == '484613',]$total_population
c_2016[c_2016[,"id"] == '484613',]$total_population

# relation of undergraduate and graduate population in 2016
# undergraduate population
data_2016 = data[data[,"academic_year"] == '2016',]
cor(data_2016$size,data_2016$grad_students,use = "complete.obs")    # 0.6960052

data_2015 = data[data[,"academic_year"] == '2015',]
cor(data_2015$size,data_2015$grad_students,use = "complete.obs")    #0.6979115

data_2014 = data[data[,"academic_year"] == '2014',]
cor(data_2014$size,data_2014$grad_students,use = "complete.obs")    # 0.712125

data_2013 = data[data[,"academic_year"] == '2013',]
cor(data_2013$size,data_2013$grad_students,use = "complete.obs")    # 0.7105582

data_2012 = data[data[,"academic_year"] == '2012',]
cor(data_2012$size,data_2012$grad_students,use = "complete.obs")    # 0.7115982

Undergraduate_Population = data_2016$size
Graduate_Population = data_2016$grad_students
hist1 = ggplot(data_2016, aes(Undergraduate_Population ,Graduate_Population)) + geom_point()
+ geom_smooth(method='lm') +
   labs(title="Relationship between Undergraduate and Graduate Populations in 2016")+
   theme(plot.title=element_text(hjust=0.5,size = 10)) + annotate("text",x=60000,y=20000,label
="Cor = 0.6960052")

hist2 = ggplot(data_2015, aes(data_2015$size,data_2015$grad_students)) + geom_point() +
geom_smooth(method='lm') +
   labs(title="Relationship between Undergraduate and Graduate Populations in 2015",
x="Undergraduate_Population", y="Graduate_Population")+
   theme(plot.title=element_text(hjust=0.5,size = 10)) + annotate("text",x=100000,y=18000,label
="Cor = 0.6979115")

hist3 = ggplot(data_2014, aes(data_2014$size,data_2014$grad_students)) + geom_point() +
geom_smooth(method='lm') +
   labs(title="Relationship between Undergraduate and Graduate Populations in 2014",
x="Undergraduate_Population", y="Graduate_Population")+
```

```
  theme(plot.title=element_text(hjust=0.5,size = 10)) + annotate("text",x=100000,y=18000,label
="Cor = 0.712125")

hist4 = ggplot(data_2013, aes(data_2013$size,data_2013$grad_students)) + geom_point() +
geom_smooth(method='lm') +
  labs(title="Relationship between Undergraduate and Graduate Populations in 2013",
x="Undergraduate_Population", y="Graduate_Population")+
  theme(plot.title=element_text(hjust=0.5,size = 10)) + annotate("text",x=100000,y=18000,label
="Cor = 0.7105582")

hist5 = ggplot(data_2012, aes(data_2012$size,data_2012$grad_students)) + geom_point() +
geom_smooth(method='lm') +
  labs(title="Relationship between Undergraduate and Graduate Populations in 2012",
x="Undergraduate_Population", y="Graduate_Population")+
  theme(plot.title=element_text(hjust=0.5,size = 10)) + annotate("text",x=100000,y=18000,label
="Cor = 0.7115982")

grid.arrange(hist1,hist2,hist3,hist4,hist5,nrow=2)

##### 3
# add two loops into one
data_year = list()
Program = NULL
for (i in (1:5)){
  data_year[[i]] = data[data[,'academic_year'] == 2011+i,]
  for (j in (47:84)){
    temp<-mean(data_year[[i]][,j], na.rm = TRUE)
    temp<-as.data.frame(temp)
    temp$Features<-colnames(data_year[[i]])[j]
    temp$Year <- 2011+i
    Program<-rbind(Program,temp)
  }
}

Program$temp = round(Program$temp,5)
names(Program)[names(Program) == "temp"] = "Mean_of_program_percentages"

######### Are there any program percentages that show patterns different from the others?
data_year = list()
for (i in (1:5)){
  data_year[[i]] = Program[Program[,'Year'] == 2011+i,]
}
Pro = list()
for (i in (1:5)){
```

```
   Pro[[i]] = ggplot(data_year[[i]]) +
      geom_point( aes(x=Features, y=Mean_of_program_percentages)) +
      labs(title = paste("The Mean of Program Percentages in", 2011 + i), x="Features", y="Mean
of Program Percentages")+
      theme(plot.title=element_text(hjust=0.5,size = 10)) + theme(axis.text.x = element_blank())
}
grid.arrange(Pro[[1]],Pro[[2]],Pro[[3]],Pro[[4]],Pro[[5]],nrow=2)

library(dplyr)
# the most/least popular programes
#       least_popular = Program[order(Program$Year,Program$Mean_of_program_percentages),]
#       most_popular = Program[order(Program$Year,-Program$Mean_of_program_percentages),]

# the most/least 5 popular programes
most_popular_2 = Program %>% group_by(Year) %>% top_n(5, Mean_of_program_percentages)
least_popular_2 = Program %>% group_by(Year) %>% top_n(-5, Mean_of_program_percentages)

###################### loop to plot 5 graphs of the most popular program
data_year = list()
for (i in (1:5)){
   data_year[[i]] = most_popular_2[most_popular_2['Year'] == 2011+i,]
}
Pro = list()
for (i in (1:5)){
   Pro[[i]] = ggplot(data_year[[i]]) +
      geom_bar( aes(x=Features, y=Mean_of_program_percentages,fill = Features),stat="identity")
+ scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9","paleturquoise3","rosybrown2"))
+
      guides(fill = "none") +
      labs(title = paste("The Most Popular Program in", 2011 + i), x="", y="Mean of Program
Percentages")+
      theme(plot.title=element_text(hjust=0.5,size = 10)) + theme(axis.text.x =
element_text(angle = 30))+
      geom_text(aes(x=Features, y=Mean_of_program_percentages,label =
Mean_of_program_percentages , vjust = -0.1, hjust = 0.5),size = 4)+
      scale_x_discrete(breaks=c("program_percentage.personal_culinary",
"program_percentage.humanities",
"program_percentage.visual_performing","program_percentage.health","program_percentage.b
usiness_marketing"),
                       labels=c("personal culinary", "humanities", "visual
performing","health","marketing"))
}
grid.arrange(Pro[[1]],Pro[[2]],Pro[[3]],Pro[[4]],Pro[[5]],nrow=2)
```

```
#################### loop to plot 5 graphs of the least popular program
data_year = list()
for (i in (1:5)){
  data_year[[i]] = least_popular_2[least_popular_2[,'Year'] == 2011+i,]
}
Pro = list()
for (i in (1:5)){
  Pro[[i]] = ggplot(data_year[[i]]) +
    geom_bar(aes(Features, Mean_of_program_percentages,fill = Features),stat="identity") +
scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9","paleturquoise3","rosybrown2")) +
    guides(fill = "none") + theme(axis.text.x = element_text(angle = 30))+
    labs(title = paste("The least Popular Program in", 2011 + i), x="", y="Mean of Program
Percentages")+
    theme(plot.title=element_text(hjust=0.5,size = 10))+
    geom_text(aes(Features, Mean_of_program_percentages,label =
Mean_of_program_percentages , vjust = -0.1, hjust = 0.5),size = 4)+
    scale_x_discrete(breaks=c("program_percentage.architecture",
"program_percentage.ethnic_cultural_gender",
"program_percentage.library","program_percentage.military","program_percentage.science_tec
hnology"),
                     labels=c("architecture", "ethnic cultural gender",
"library","military","science technology"))
}
grid.arrange(Pro[[1]],Pro[[2]],Pro[[3]],Pro[[4]],Pro[[5]],nrow=2)

#### try one plot
least_popular_2014 =  least_popular_2[least_popular_2[,'Year'] == "2014",]
ggplot(least_popular_2014, aes(x=Features, y=Mean_of_program_percentages,fill = Features))
+guides(fill = "none") +
  geom_bar(stat="identity") + scale_fill_manual(values=c("#999999", "#E69F00",
"#56B4E9","paleturquoise3","rosybrown2")) +
   geom_text(aes(label = Mean_of_program_percentages , vjust = -0.4, hjust = 0.5),size = 4)+
  scale_x_discrete(breaks=c("program_percentage.architecture",
"program_percentage.ethnic_cultural_gender",
"program_percentage.library","program_percentage.military","program_percentage.science_tec
hnology"),
                   labels=c("architecture", "ethnic_cultural_gender",
"library","military","science_technology"))

################################## 4
tuition = data[,c("id","state","tuition.in_state","tuition.out_of_state","academic_year")]

# add "tuition.in_state","tuition.out_of_state" into "total_tuition"
```

```
tuition$tuition.in_state[is.na(tuition$tuition.in_state)] = 0
tuition$tuition.out_of_state[is.na(tuition$tuition.out_of_state)] = 0

tuition$total_tuition = tuition$tuition.in_state + tuition$tuition.out_of_state
tuition$total_tuition[tuition$total_tuition == 0] = NA

var_tuition = tuition %>% group_by(state) %>%
   mutate(Var_tuition = var(total_tuition,na.rm = TRUE))
var_tuition = var_tuition[,c("state","Var_tuition","academic_year")]
var_tuition = unique(var_tuition)
var_tuition_most = var_tuition %>% group_by(academic_year) %>% top_n(5, Var_tuition)
var_tuition_least = var_tuition %>% group_by(academic_year) %>% top_n(-5, Var_tuition)

tuition = data[,c("id","state","tuition.in_state","tuition.out_of_state","academic_year")]
# var of tuition.in_state
var_tuition.in_state = tuition %>% group_by(state) %>%
   mutate(Var_instate = var(tuition.in_state,na.rm = TRUE))
var_tuition.in_state = var_tuition.in_state[,c("state","Var_instate","academic_year")]
var_tuition.in_state = unique(var_tuition.in_state)

# var of tuition.out_of_state
var_tuition.out_of_state = tuition %>% group_by(state) %>%
   mutate(Var_outstate = var(tuition.out_of_state,na.rm = TRUE))
var_tuition.out_of_state = var_tuition.out_of_state[,c("state","Var_outstate","academic_year")]
var_tuition.out_of_state = unique(var_tuition.out_of_state)

# the most/least 5 var tuition.in_state
var_tuition.in_state_most = var_tuition.in_state %>% group_by(academic_year) %>% top_n(5,
Var_instate)
var_tuition.in_state_least = var_tuition.in_state %>% group_by(academic_year) %>% top_n(-5,
Var_instate)

var_tuition.out_of_state_most = var_tuition.out_of_state %>% group_by(academic_year) %>%
top_n(5, Var_outstate)
var_tuition.out_of_state_least = var_tuition.out_of_state %>% group_by(academic_year) %>%
top_n(-5, Var_outstate)


#relationship between the number of universities in a state and total tuition
tuition
mean_tuition = tuition %>% group_by(state) %>%
   mutate( mean_total_tuition = mean(total_tuition,na.rm = TRUE))
mean_tuition = mean_tuition[,c("state","mean_total_tuition")]
mean_tuition = unique(mean_tuition)
```

```
mean_tuition = mean_tuition[order(mean_tuition$mean_total_tuition),]
table(tuition$state)

##### 5
race =
data[,c("id","name","demographics.race_ethnicity.white","demographics.race_ethnicity.black",
"demographics.race_ethnicity.hispanic","demographics.race_ethnicity.asian","demographics.race
_ethnicity.aian","demographics.race_ethnicity.nhpi","demographics.race_ethnicity.unknown")]
#Find the college which has the most variance of the percentage of race
apply(race[,3:9],1,var)

race$name[which.max(apply(race[,3:9],1,var))]    # "Marsha Kay Beauty College"

# replace NA to 0
# race[is.na(race)] = 0

# use men/women to get the ratio to measure the diversity of demographics
wm =    data[,c("id","name","demographics.men","demographics.women")]
wm$ratio = wm$demographics.men/wm$demographics.women
wm = wm[order(wm$ratio),]
# if the ratio are close to 1, the school would have more diversity.

### 6 (a)
data_ca = data[data["state"] == 'CA',]
data_year = list()
Program = NULL
for (i in (1:5)){
   data_year[[i]] = data_ca[data_ca['academic_year'] == 2011+i,]
   for (j in (47:84)){
      temp<-mean(data_year[[i]][,j], na.rm = TRUE)
      temp<-as.data.frame(temp)
      temp$Features<-colnames(data_year[[i]])[j]
      temp$Year <- 2011+i
      Program<-rbind(Program,temp)
   }
}

Program$temp = round(Program$temp,5)
names(Program)[names(Program) == "temp"] = "Mean_of_program_percentages"

# the most/least 5 popular programes
most_popular_ca = Program %>% group_by(Year) %>% top_n(5, Mean_of_program_percentages)
least_popular_ca = Program %>% group_by(Year) %>% top_n(-5,
Mean_of_program_percentages)
```

```
##################### loop to plot 5 graphs of the most popular program
data_year = list()
for (i in (1:5)){
  data_year[[i]] = most_popular_ca[most_popular_ca[,'Year'] == 2011+i,]
}
Pro = list()
for (i in (1:5)){
  Pro[[i]] = ggplot(data_year[[i]]) +geom_bar( aes(x=Features, y=Mean_of_program_percentages,fill = Features),stat="identity") + scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9","paleturquoise3","rosybrown2")) +guides(fill = "none") +labs(title = paste("The Most Popular Program in California, in", 2011 + i), x="", y="Mean of Program Percentages")+ theme(plot.title=element_text(hjust=0.5,size = 10)) + theme(axis.text.x = element_text(angle = 30))+ geom_text(aes(x=Features, y=Mean_of_program_percentages,label = Mean_of_program_percentages , vjust = 0, hjust = 0.5),size = 4)
}
grid.arrange(Pro[[1]],Pro[[2]],Pro[[3]],Pro[[4]],Pro[[5]],nrow=2)

##################### loop to plot 5 graphs of the least popular program
data_year = list()
for (i in (1:5)){
  data_year[[i]] = least_popular_ca[least_popular_ca[,'Year'] == 2011+i,]
}
Pro = list()
for (i in (1:5)){
  Pro[[i]] = ggplot(data_year[[i]]) +
    geom_bar(aes(Features, Mean_of_program_percentages,fill = Features),stat="identity") + scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9","paleturquoise3","rosybrown2")) +
    guides(fill = "none") + theme(axis.text.x = element_text(angle = 30))+
    labs(title = paste("The least Popular Program in", 2011 + i), x="", y="Mean of Program Percentages")+
    theme(plot.title=element_text(hjust=0.5,size = 10))+
    geom_text(aes(Features, Mean_of_program_percentages,label = Mean_of_program_percentages , vjust = 0, hjust = 0.5),size = 4)
}
grid.arrange(Pro[[1]],Pro[[2]],Pro[[3]],Pro[[4]],Pro[[5]],nrow=2)

# 6 (b)
ca_2016 = data_ca[data_ca[,'academic_year'] == "2016",]
data_ca$tuition.in_state[which(data_ca$name=="University of California-Davis")]    # 14046
data_ca$tuition.out_of_state[which(data_ca$name=="University of California-Davis")] # 40728

# get UCD row
```

```
which(ca_2016$name=="University of California-Davis")
ca_2016[50,]

hist1 = ggplot(ca_2016, aes(x = as.factor(id), y = tuition.in_state )) + geom_point()+
  theme(axis.ticks.x = element_blank() )+ theme(plot.title=element_text(hjust=0.5,size =20))+
  labs(title="In State Tuition for the California Universities: 2016",x="Universities",y="In State Tuition")+theme(axis.text.x = element_blank() )+
  geom_point(data = ca_2016[50,], aes(x = as.factor(id), y = tuition.in_state ), colour = "#E69F00", size = 5) +annotate("text",x=90,y=17000,label ="University of California-Davis in state tuition : 14046 ",size = 6)

hist2 = ggplot(ca_2016, aes(x = as.factor(id), y = tuition.out_of_state )) + geom_point()+
  theme(axis.ticks.x = element_blank() )+ theme(plot.title=element_text(hjust=0.5,size =20))+
  labs(title="Out of State Tuition for the California Universities: 2016",x="Universities",y="Out of State Tuition")+
  theme(axis.text.x = element_blank() ) +
  geom_point(data = ca_2016[50,], aes(x = as.factor(id), y = tuition.out_of_state ), colour = "#E69F00", size = 5)+
  annotate("text",x=90,y=45000,label ="University of California-Davis
           out of state tuition : 40728 ",size = 6)

mean_instate_tuition = mean(ca_2016$tuition.in_state,na.rm = TRUE)      # 14625
mean_outofstate_tuition = mean(ca_2016$tuition.out_of_state,na.rm = TRUE) # 18376

# 8 (a)
data_Q8 = data[,c("id","name","tuition.out_of_state","demographics.race_ethnicity.white","demographics.race_ethnicity.black",
"demographics.race_ethnicity.hispanic","demographics.race_ethnicity.asian","demographics.race_ethnicity.aian",
"demographics.race_ethnicity.nhpi","demographics.race_ethnicity.unknown","academic_year")]
data_Q8$diversity_race = apply(data_Q8[,4:10],1,var)
data_Q = data_Q8[,c("id","name","tuition.out_of_state","diversity_race",'academic_year')]

data_year = list()
for (i in (1:5)){
  data_year[[i]] = data_Q[data_Q[,'academic_year'] == 2011+i,]
}
Pro = list()
for (i in (1:5)){
  Pro[[i]] = ggplot(data_year[[i]]) +
    geom_point( aes(x=diversity_race, y=tuition.out_of_state)) + geom_smooth(aes(x=diversity_race, y=tuition.out_of_state),method='lm')+
    labs(title = paste("Relationship between Out of State Tuition and Race Diversity, in", 2011 +
```

```
i), x="Race Diversity", y="Out of State Tuition")+
      theme(plot.title=element_text(hjust=0.5,size = 10)) + theme(axis.text.x = element_blank())
}
grid.arrange(Pro[[1]],Pro[[2]],Pro[[3]],Pro[[4]],Pro[[5]],nrow=2)

# calculate the cor of race diversity and out of state tuition
cor(data_year[[1]]$diversity_race, data_year[[1]]$tuition.out_of_state,use = "complete.obs") #
-0.16720
cor(data_year[[2]]$diversity_race, data_year[[2]]$tuition.out_of_state,use = "complete.obs") #
-0.17135
cor(data_year[[3]]$diversity_race, data_year[[3]]$tuition.out_of_state,use = "complete.obs") #
-0.18209
cor(data_year[[4]]$diversity_race, data_year[[4]]$tuition.out_of_state,use = "complete.obs") #
-0.19415
cor(data_year[[5]]$diversity_race, data_year[[5]]$tuition.out_of_state,use = "complete.obs") #
-0.20903

# (b)
data_ownership = data[,c("id","name","state","tuition.in_state","ownership",'academic_year')]
ownership_2016 = data_ownership[data_ownership["academic_year"] == '2016',]
ownership_ca_2016 = ownership_2016[ownership_2016["state"] == 'CA',]
ownership_ca_2016 = na.omit(ownership_ca_2016)
id = as.factor(ownership_ca_2016$id)
ownership = as.factor(ownership_ca_2016$ownership)

ggplot(ownership_ca_2016,aes(x=reorder(id,-tuition.in_state),y=tuition.in_state,fill=ownership))
+geom_bar(stat="identity")+labs(title="In State Tuition Distribution in California Schools by
Ownership in 2016",x="Schools",y="In State Tuition")+
   theme(plot.title=element_text(hjust=0.5,size=20))+theme(axis.text.x = element_blank() )+
   scale_fill_manual(values = c("#8FBC94", "lightskyblue","darkorange"))+theme(axis.ticks.x =
element_blank())+
   geom_hline(yintercept=14625,color="black",size=1.5)+
   annotate('text',x=230,y=17000,label="California Satewide Average 14625",size=5)+
   theme(legend.position = c(1, 1),legend.justification = c(1,1))

# calculate the California statewide average tuition-in-state
mean(ownership_ca_2016$tuition.in_state,na.rm = TRUE)
```