**141A HW2**

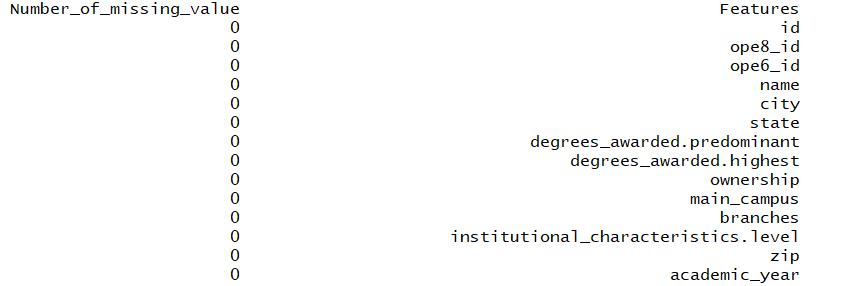
**Man Pan**

**Student ID: 914656278**

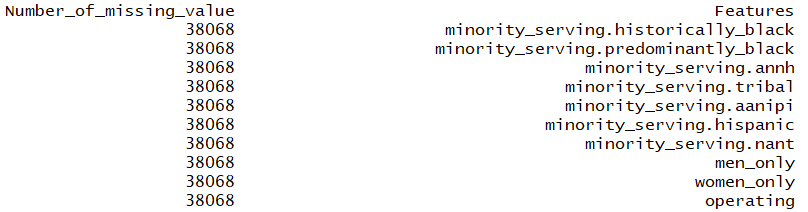
**Email: [manpan@ucdavis.edu](mailto:manpan@ucdavis.edu)**

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:



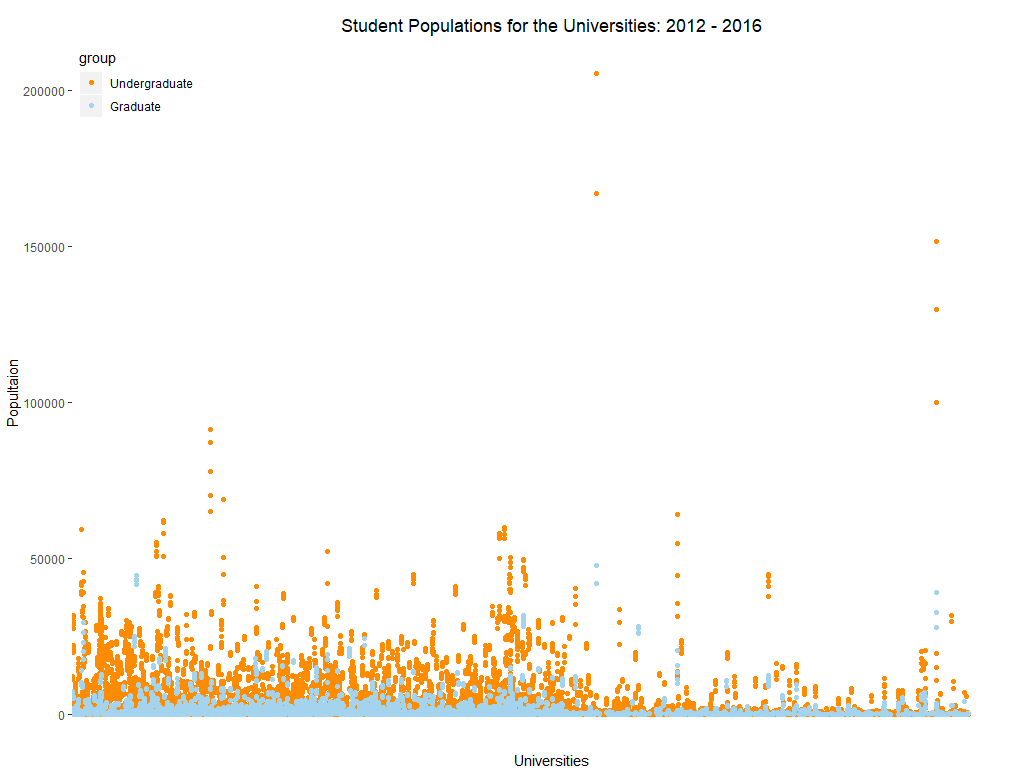
features have the most missing values:



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.

1. **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;

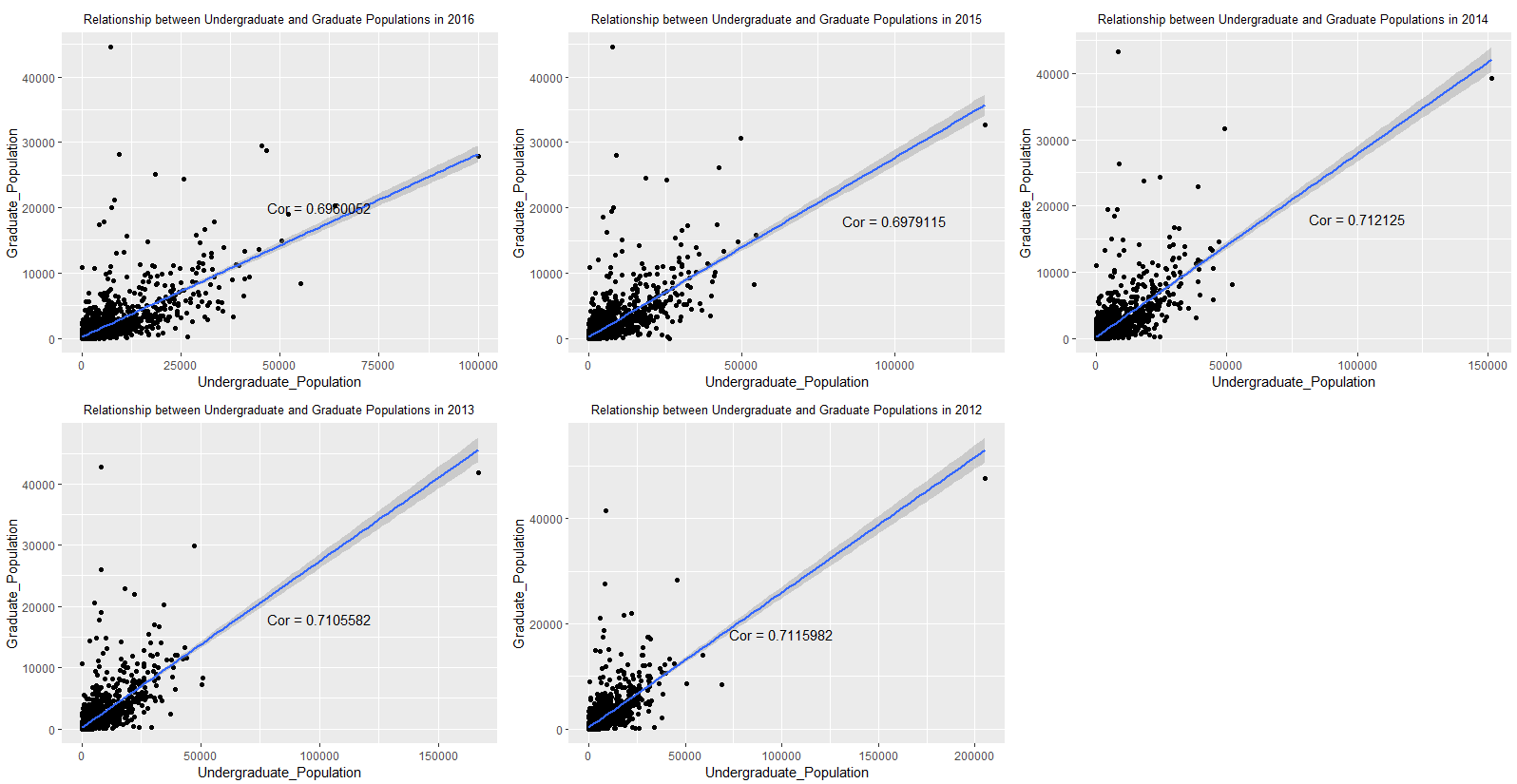
In 2014, University of Phoenix-Arizona has unusual population. The total population is 190745;

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.

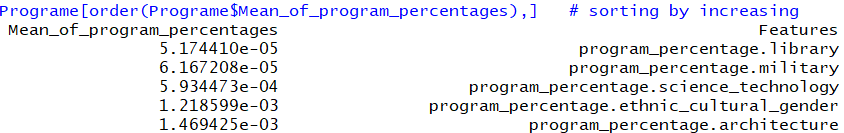
Finally, I select the population data in 2012-2016 to explore the relation between Undergraduate and graduate populations.

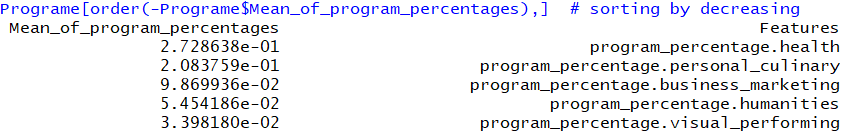


From the plot, we can find that the correlation is 0.6960052 (2016), 0.6979115 (2015), 0.712125 (2014), 0.7105582 (2013), 0.7115982 (2012), which means undergraduate and graduate population are positively correlated. I think there is no exception to the relationship. They just have different coefficient,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

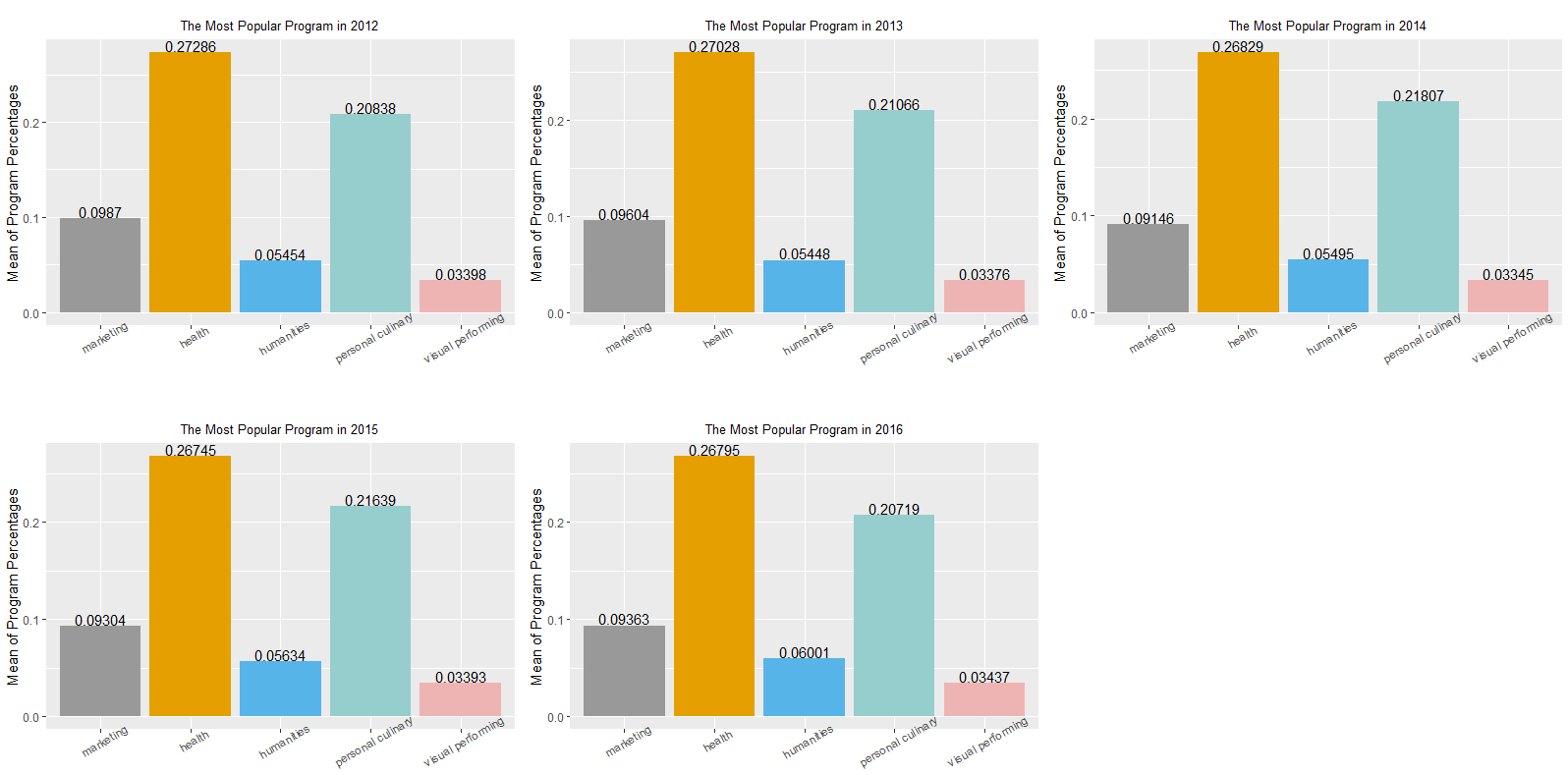
1. **Explore the program percentages for the universities. What programs are the most popular? What programs are the least popular? Are there any program percentages that show patterns different from the others?**

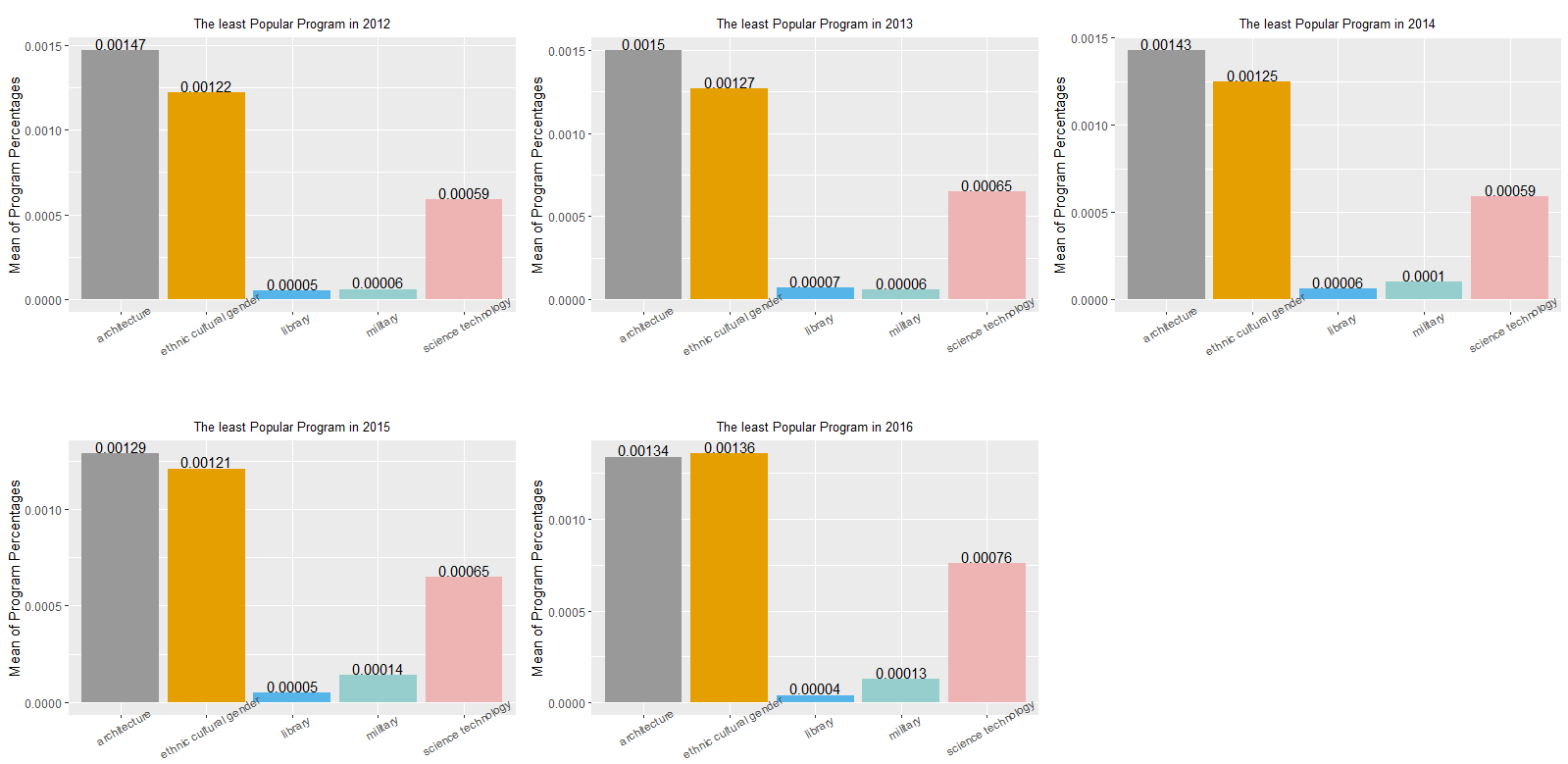
I use the mean of the program percentage to make sure which programs are the most popular and the least popular. Use the 2012 year data as an example:





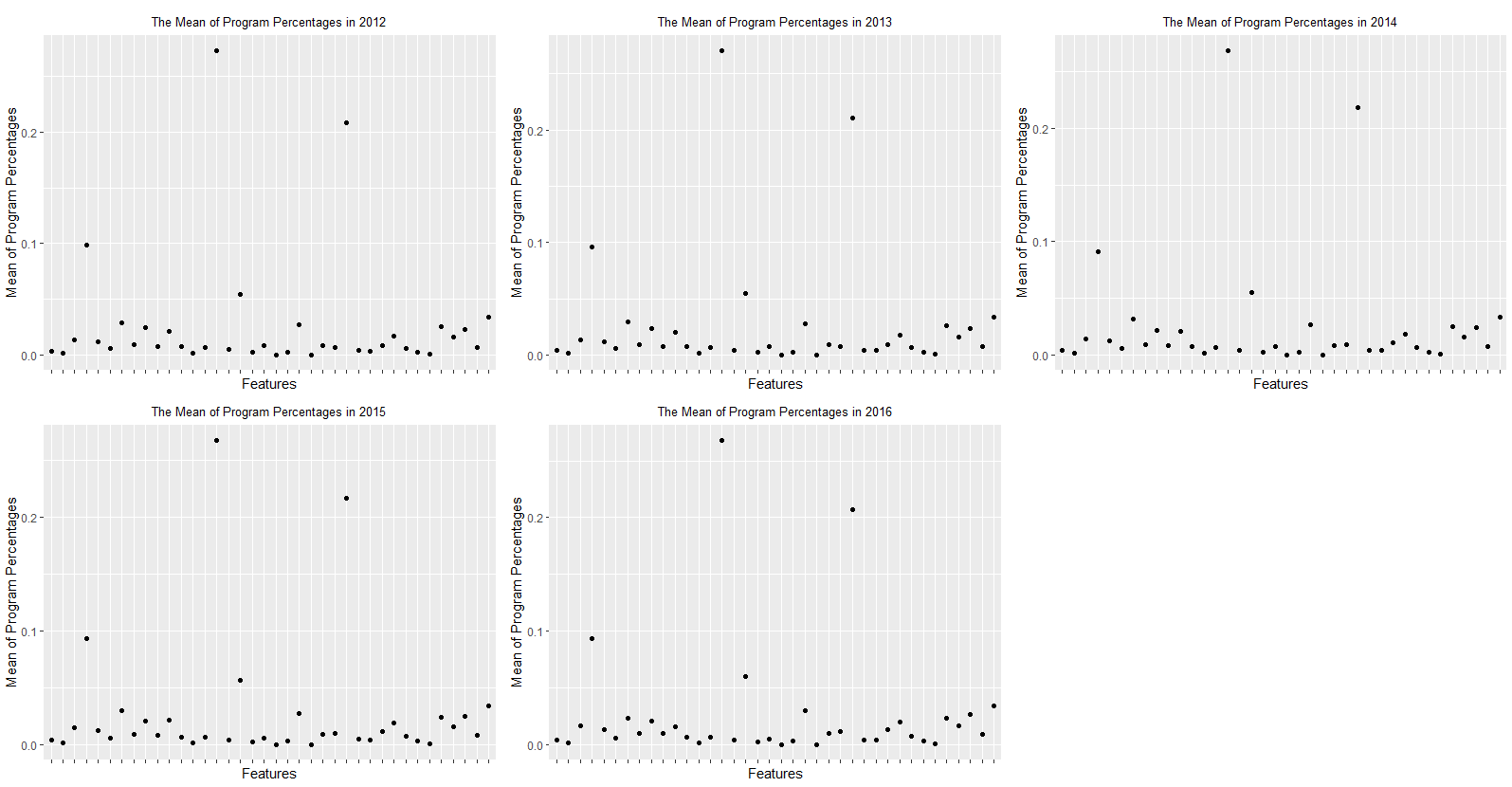
Then, I plot 5 graphs in each year to present the mean of program percentages (just select the top 5 most/least popular programs).





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.

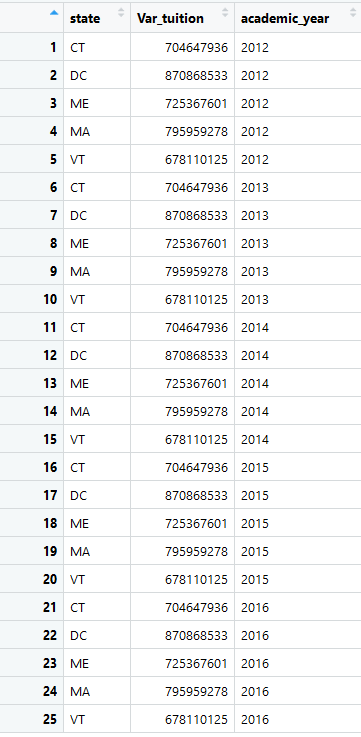
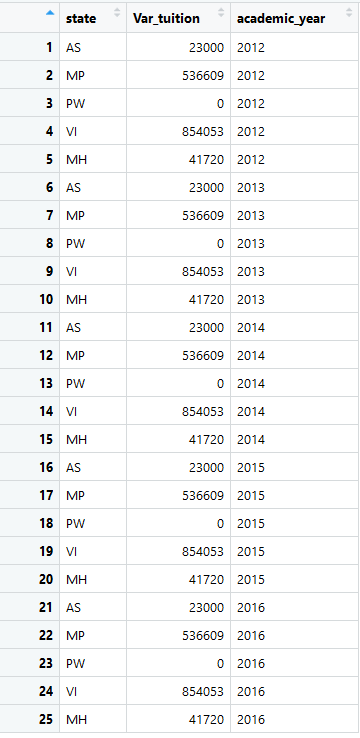
1. **How does tuition vary across different states? Is there a relationship between the number of universities in a state and tuition? Do these characteristics differ for in-state tuition and out-of-state tuition?**

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): The 5 least diverse state(total tuition):

2012 -2016 2012 - 2016

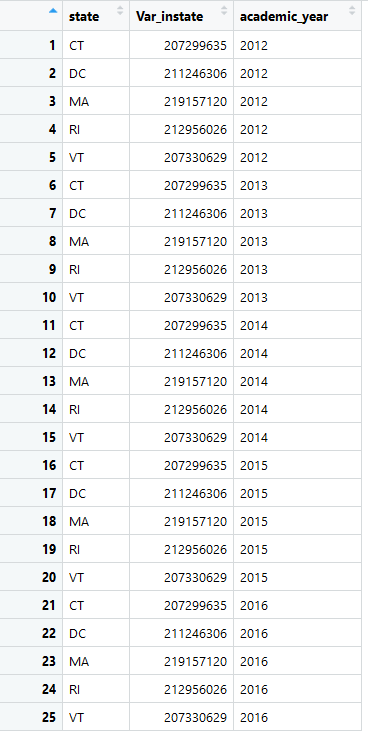
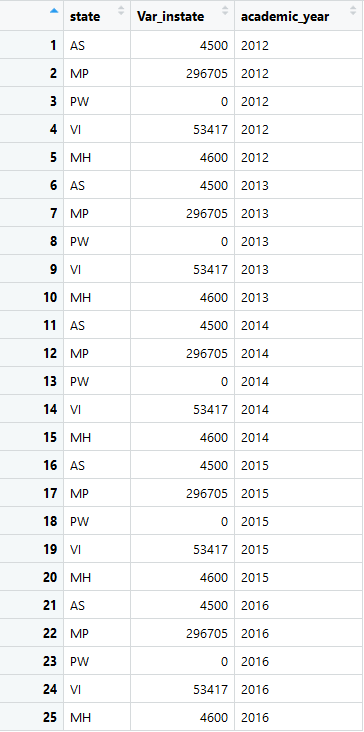
We can find DC, MA, ME, CT, VT states have the most diverse in total tuition;

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): The 5 least diverse state(in state tuition):

2012 -2016 2012 - 2016

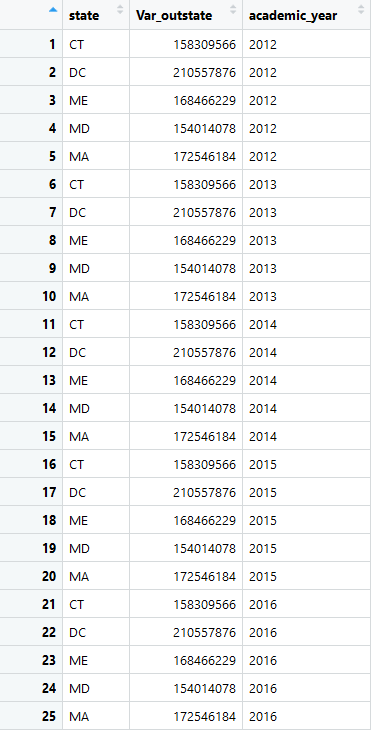
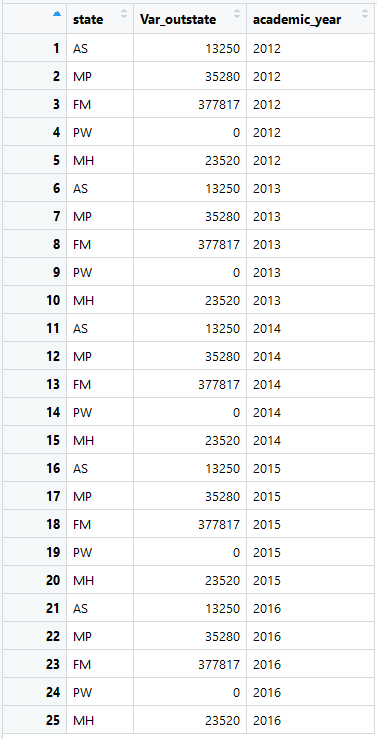


For in state tuition, we find CT, DC, MA, RI, VT states have the most diverse;

AS, MP, PW, VI ,MH states have the least diverse.

The 5 most diverse state (out of state tuition): The 5 least diverse state(out of state tuition):

2012 -2016 2012 - 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.

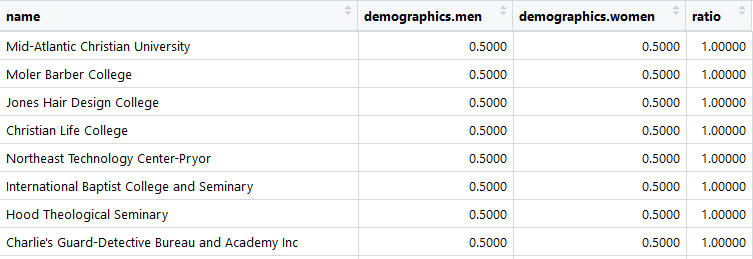
**Relationship between the number of universities in a state and tuition**

(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.

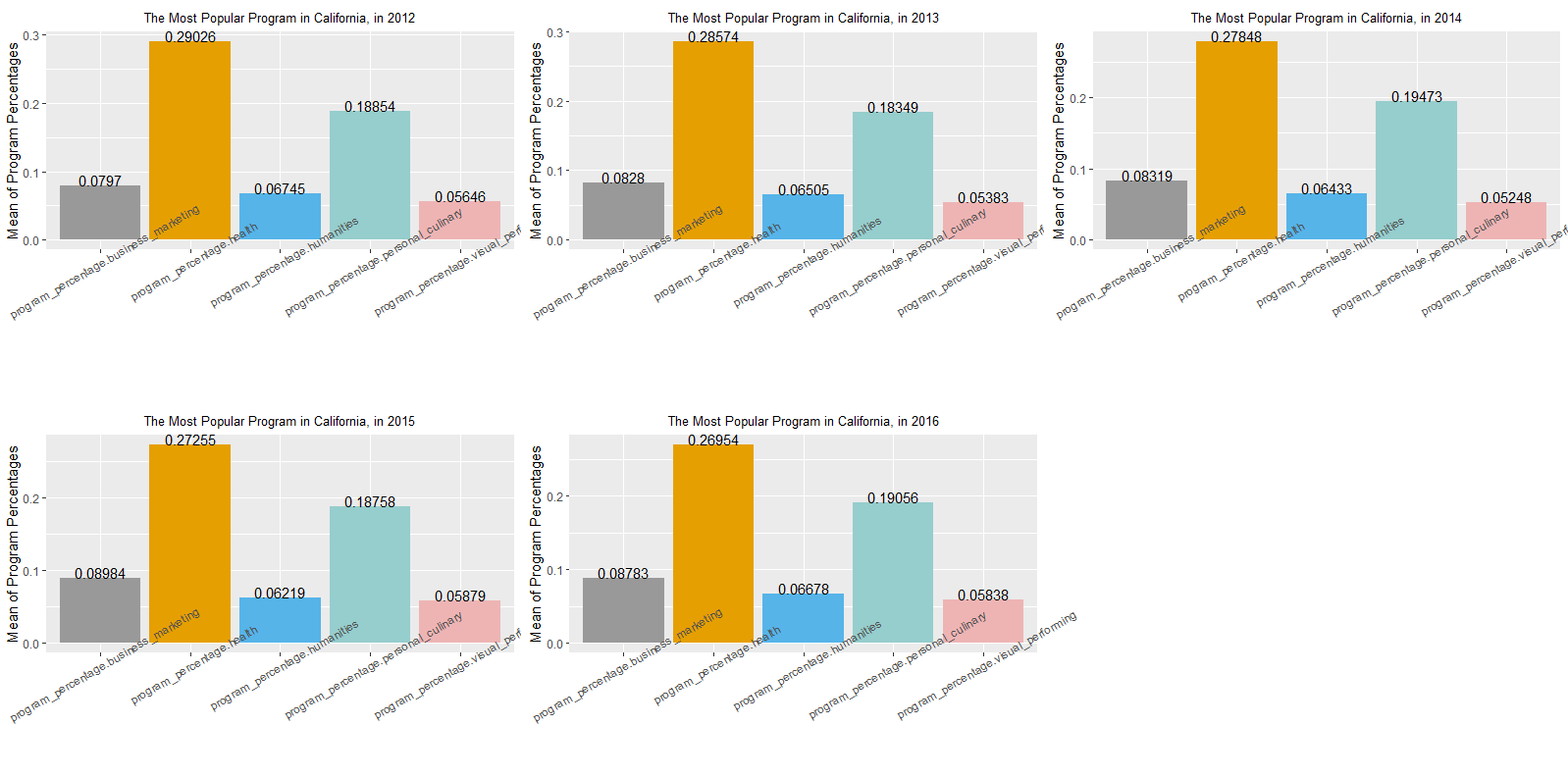
1. **Which colleges have the most diverse demographics? Make sure to explain how you measured “diversity” for this problem, in addition to discussing your conclusions .**

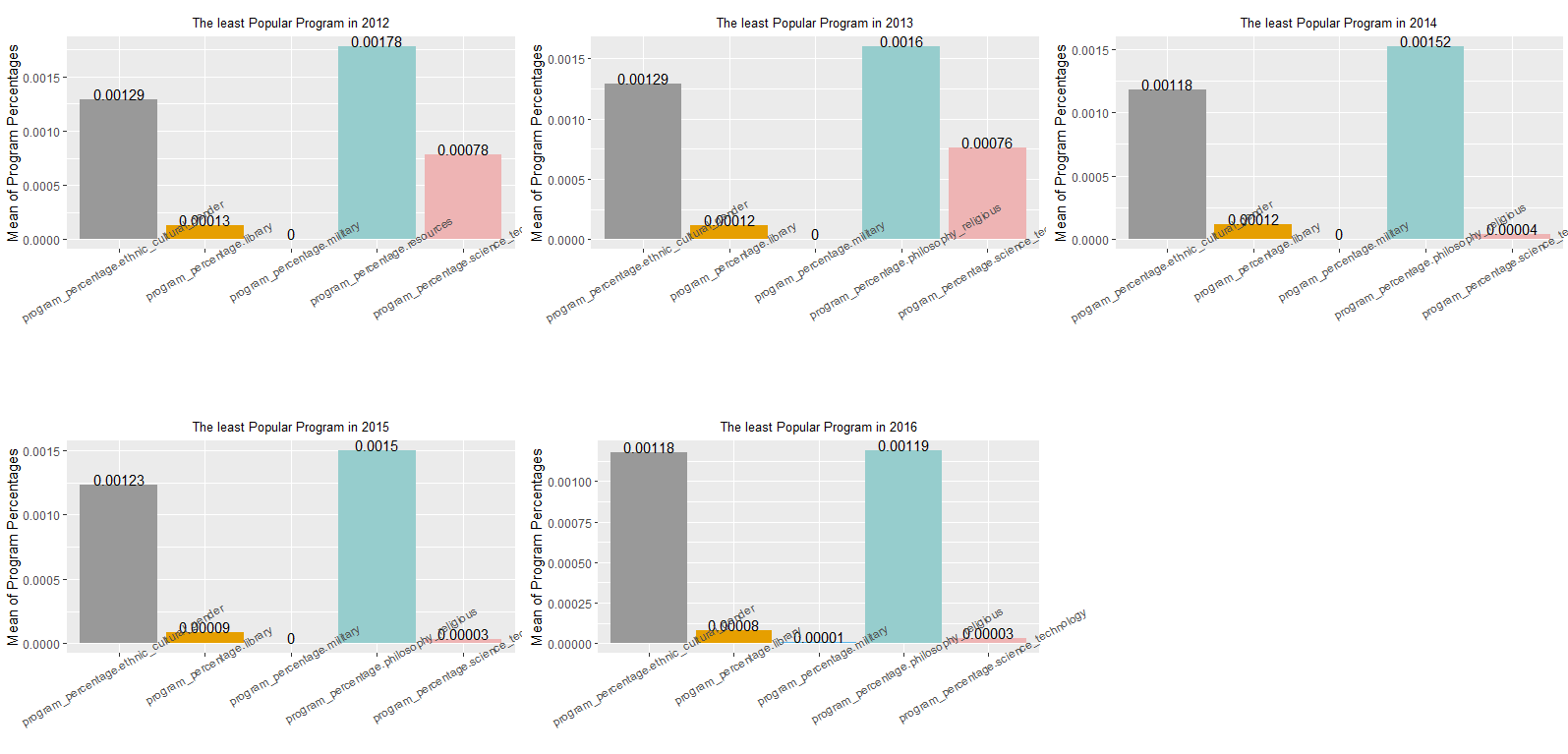
Firstly, I choose "demographics.men" and "demographics.women", then, using demographics.men /demographics.women (ratio) to as the first criterion to measure “diversity”. If the ratios are close to or equal to 1, the schools would have the most diverse demographics. I choose some schools that have the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



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).

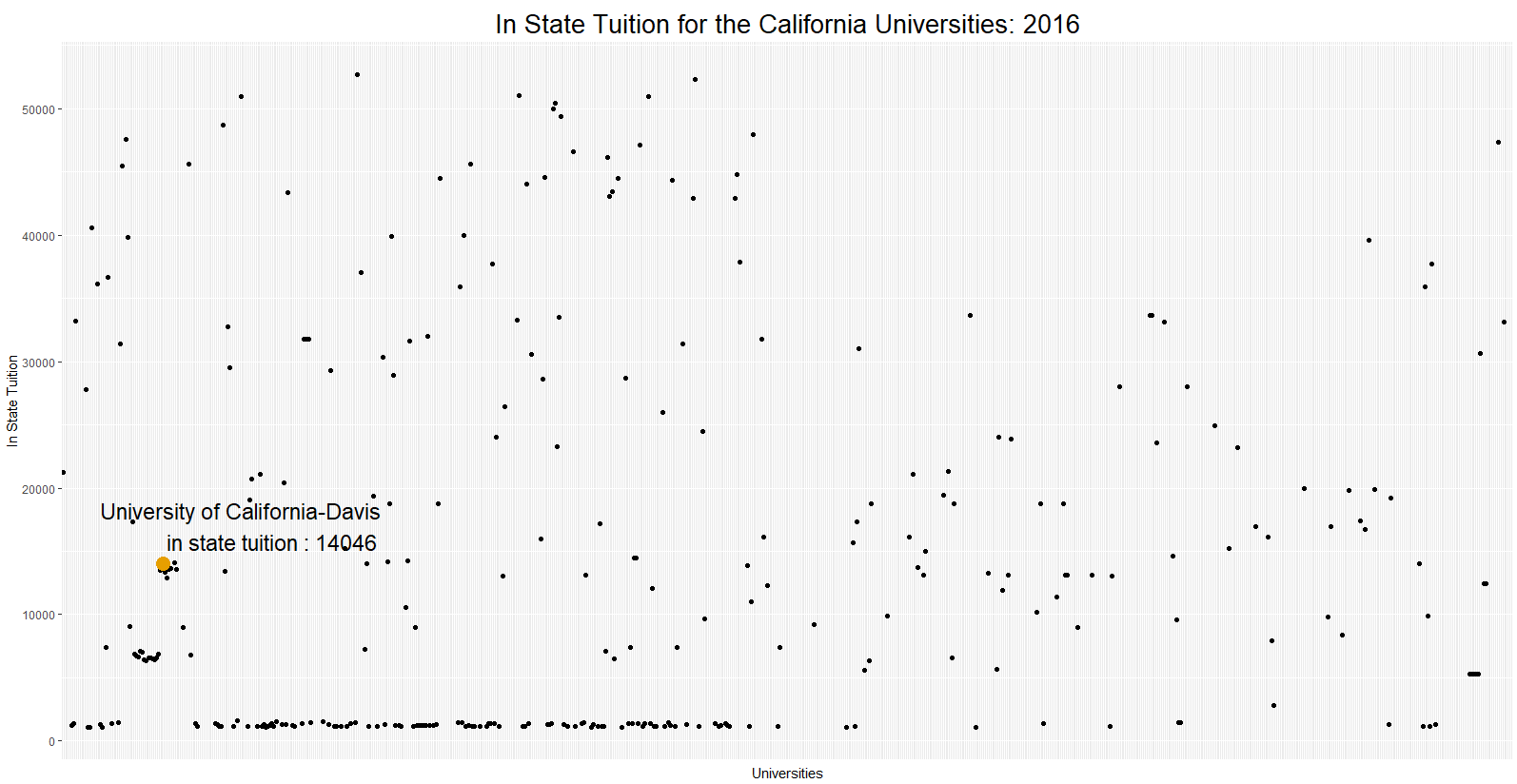
**6.(a)Compare program percentage in California colleges.( Find the most/least popular program in California)**

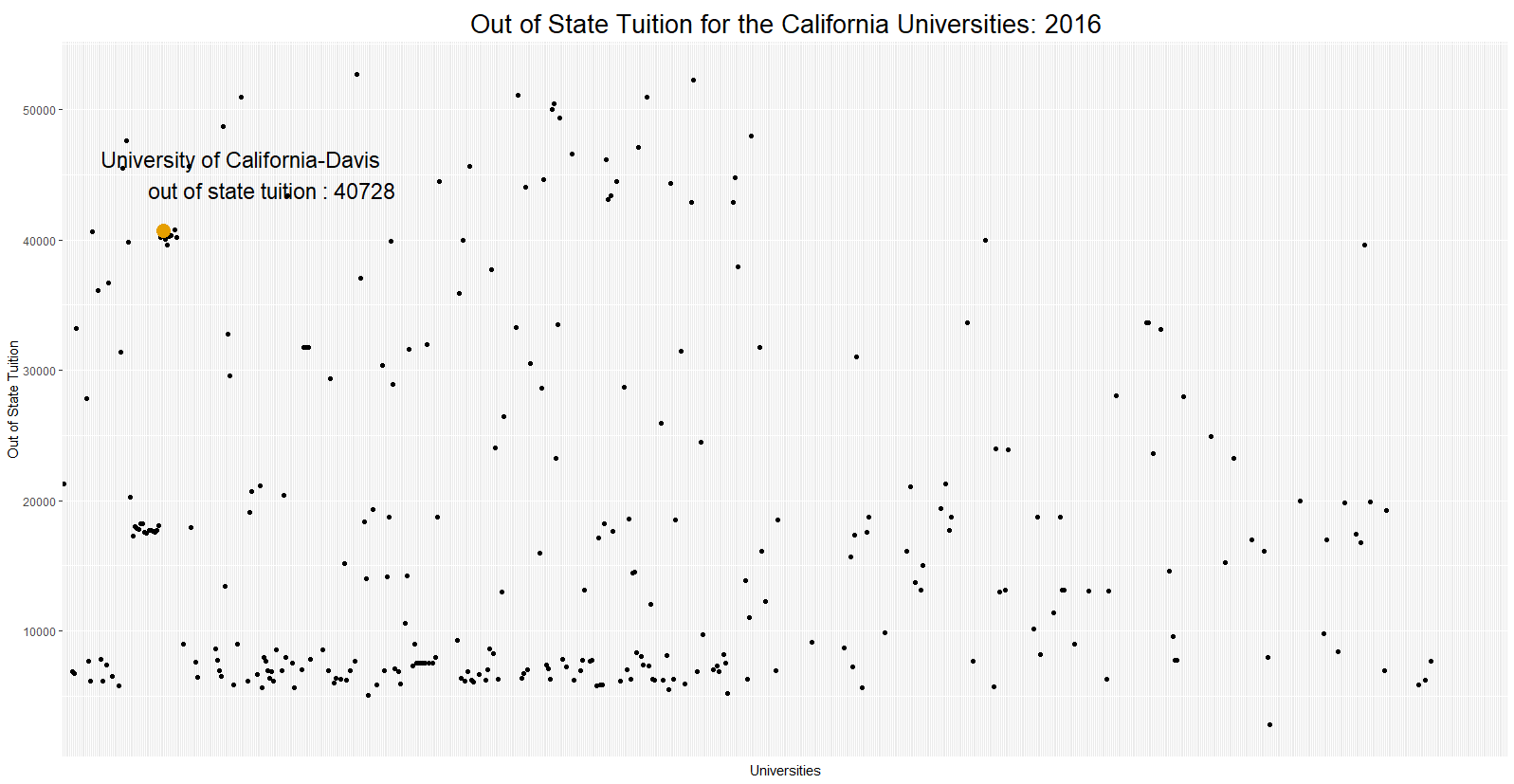




Conclusion：In California, the most popular programs are humanities, visual\_performing, health, business\_marketing , personal\_culinary; the least popular programs are resources, ethnic\_cultural\_gender, library, military, science\_technology, philosophy\_religious, science\_technology.

**(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, ethnic\_cultural\_gender, library, military, science\_technology, philosophy\_religious, science\_technology, while the least popular programs in US are Library, military, science\_technology,ethnic\_cultural\_gender, architecture. They are different. And different year has different least popular programs.

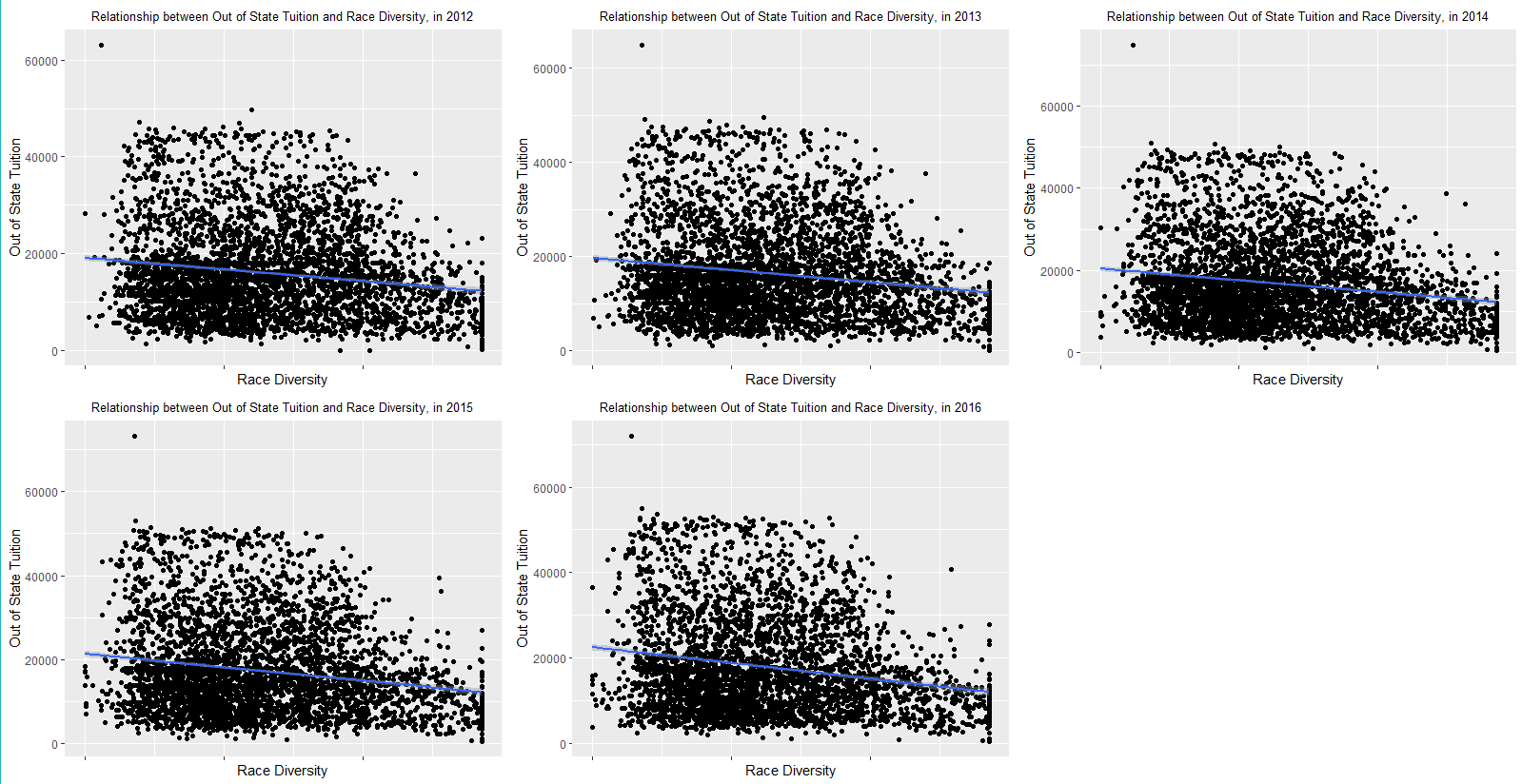
**new question:** Are the most popular programs the same in other State? Are there some schools that do not have these popular programs?

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?

1. **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.**
2. **Is there a relationship between the diversity of race and out of state tuition ?**

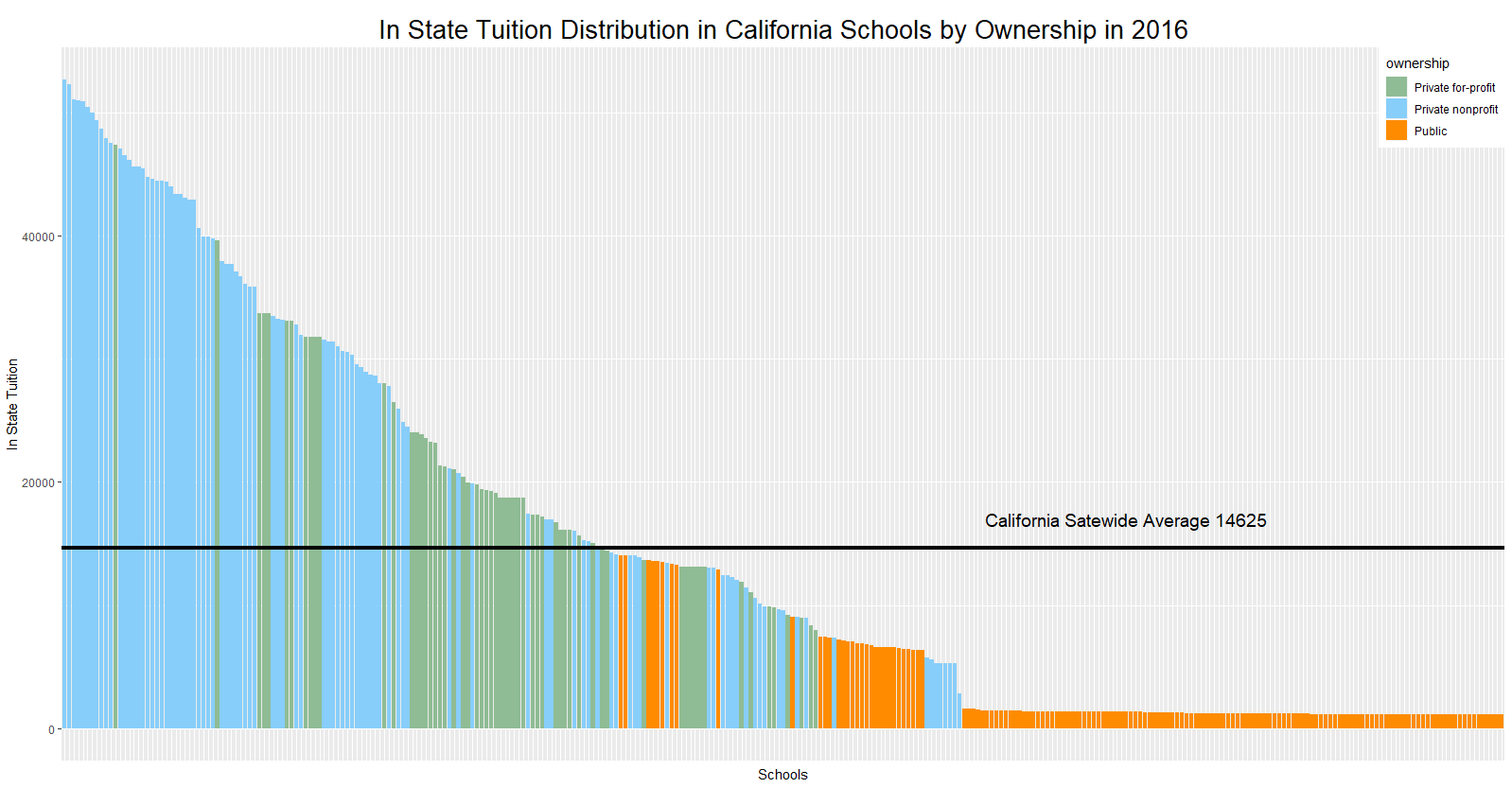
I find UCD out of state tuition is much higher than other universities’, while UCD in state tuition is lower. And I know, there are so many national students in UCD. So I want to explore the relationship between race diversity and out of state tuition.



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.

1. **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.business\_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\_technology"),

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\_technology"),

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)