Huifen Zhou’s Project

Analysis of Factors Affecting University Ranking

1. Introduction

University ranking plays a key role for students in determining which university to go. There are many factors affecting the ranking. It is important to identify the most important ones contributing to ranking. The findings could serve as a guide for university management and faculty members to improve their ranking. In this project, the best universities in the world and their geographical distribution are studied first by using R. Then regression models are developed to analyze variables used in rankings. It has been found that teaching, research, citations are the top 3 factors determining the ranking of a university.

2. DATA

A). Data source

https://www.kaggle.com/mylesoneill/world-university-rankings

B).Summary of Data

The dataset contains 13 variables named as below and has 2603 observations. The list as below

```
data.frame':                2603 obs. of  13 variables:
 $ world_rank           : chr  "1" "2" "3" "4" ...
 $ university_name       : chr  "Harvard University" "California Institute of Technology" "Massachusetts Institute
of Technology" "Stanford University" ...
 $ country              : chr  "United States of America" "United States of America" "United States of America" "Uni
ted States of America" ...
 $ teaching             : num  99.7 97.7 97.8 98.3 90.9 90.5 88.2 84.2 89.2 92.1 ...
 $ international        : chr  "72.4" "54.6" "82.3" "29.5" ...
 $ research             : num  98.7 98 91.4 98.1 95.4 94.1 93.9 99.3 94.5 89.7 ...
 $ citations            : num  98.8 99.9 99.9 99.2 99.9 94 95.1 97.8 88.3 91.5 ...
 $ income               : chr  "34.5" "83.7" "87.5" "64.3" ...
 $ total_score          : chr  "96.1" "96" "95.6" "94.3" ...
 $ num_students         : chr  "20,152" "2,243" "11,074" "15,596" ...
 $ student_staff_ratio   : num  8.9 6.9 9 7.8 8.4 11.8 11.6 16.4 11.7 4.4 ...
 $ international_students: chr  "25%" "27%" "33%" "22%" ...
 $ year                 : int  2011 2011 2011 2011 2011 2011 2011 2011 2011 2011 ...
```

C). Issues of Dataset

- Missing Value in the columns are shown on the list: world rank, income and total score.

- The column of the world_rank has unclear data such as 201-300.

• Some columns' type should be numeric but they are character variables such as total_score, num_students..

3. Using R

A) SQL

1. Top1 University from 2011 to 2016

Table 2 shows the top 1 university from 2011 to 2016. It can be learned that in 2011 Harvard University is the top 1 university. From 2012 to 2016 the California Institute of Technology is the top 1 university for 5 years.

```
  worldRank                  university_name                  country year
1         1                Harvard University United States of America 2011
2         1 California Institute of Technology United States of America 2012
3         1 California Institute of Technology United States of America 2013
4         1 California Institute of Technology United States of America 2014
5         1 California Institute of Technology United States of America 2015
6         1 California Institute of Technology United States of America 2016
```

Table 2. Top 1 University in the world from 2011 to 2016

2. Ranking of University of Cincinnati

I'd like to learn more about my own university. The highest ranking is 190th in 2011 while from 2012 to 2106 the ranking is not available as shown in Table 3 below.

```
  world_rank          university_name total_score year
1        190 University of Cincinnati        46.9 2011
2         NA University of Cincinnati          NA 2012
3         NA University of Cincinnati          NA 2013
4         NA University of Cincinnati          NA 2014
5         NA University of Cincinnati          NA 2015
6         NA University of Cincinnati          NA 2016
```

Table 3. University of Cincinnati's Rank from 2011 to 2016

B) Plot

1. The Ranking Trend of Duke, Harvard and Princeton University.

I just show 3 universities which I am interested in.

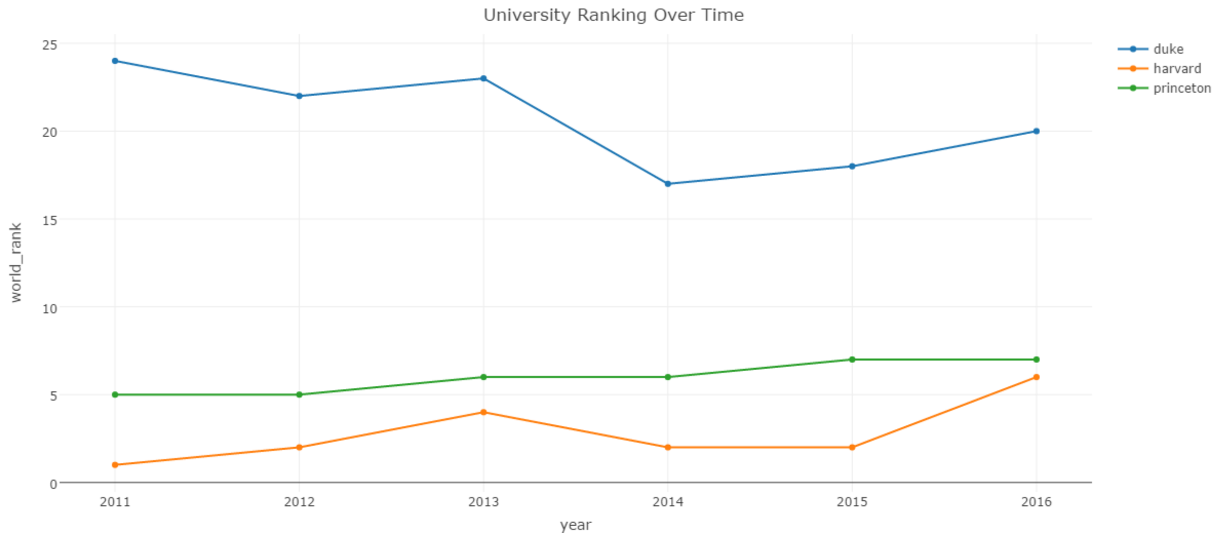


Figure 1. 3 Universities' Ranking Trend

Figure 1 shows the trend of the ranking of these 3 different universities.

The Harvard University is the best one in 2011, and its ranking changed year by year and down to No.6 in 2016.

The ranking of Princeton University is stable, which is always in the range of No.5 to No.7.

The Duke University has a gradual improvement over time, from No.24 in 2011 to No.20 in 2016.

2. The Top 10 Countries by the Number of Universities (Top 200 Universities)

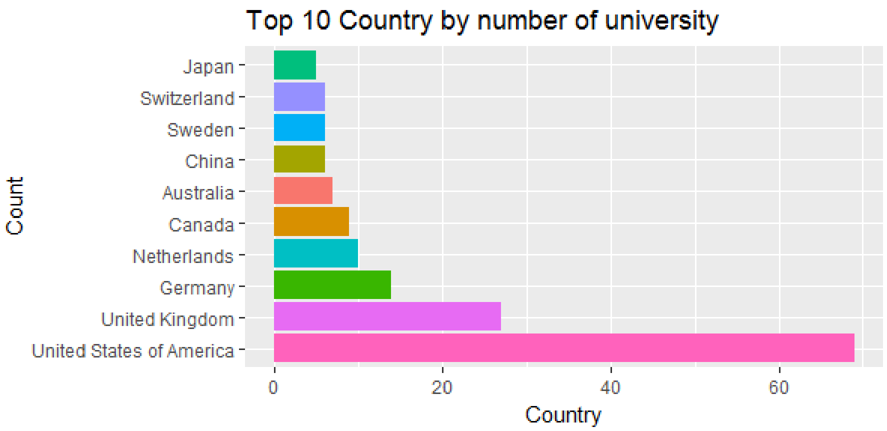


Figure 2. TOP 10 countries by number of top 200 universities

From the Figure 2, it can be learned that the USA has the most of top 200 universities and the number is near 70 universities. The UK is the No.2 country which has over than 20 universities. And other top 200 universities belong to Germany, Netherlands, Canada, Australia, China, Sweden, Switzerland and Japan. To make the results more visually clear, the map of the geographical distribution density of these countries is generated.

3. Mapping of Top 10 Counties by the Number of Top 200 Universities

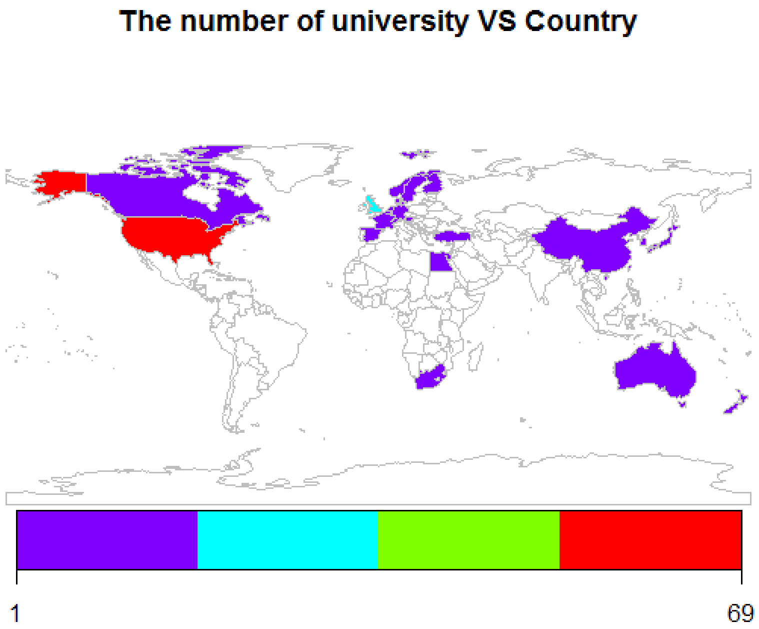


Figure 3. The Top 200 Universities' distribution

From Figure 3, it is obvious that United States has the most universities of top 200. The United Kingdom is the second country. And other universities are distributed in Canada, China, Austria, Japan and other European countries.

4. Regression

Section 1. Introduction of Data Set and Purpose of Project

a) Introduction of Data Set

To get the ranking of a university, we use scores to evaluate it. To get the scores, certain factors are used. Here I propose to use the total score as the dependent variable; teaching, research, citations and other variables as independent variables.

This study provides top 100 universities as observation in 2016. I want to find the relationship between dependent variables and independent variables.

b) Data Clean

As I mentioned in the data part, some variables should be numeric variables but instead they are character variables. Then the first step is to change the variable type.

I dropped observations which contain the missing value. Then 99 observations will be used to do the linear regression.

Section 2. Data Analysis

a) Correlation between independent variables

From table 1 in appendix, we know that there is a strong positive correlation between *teaching* and *research* (0.87), and between *international* and *international students* (0.82). As we know, *teaching* and research are the important factors to a university. However, the score of *international* depends on *international_stud.* Maybe the international will contain *internation_stud* and the P-value of the correlation between *internation_stud* and international is 0.000. We conclude that they have significant relationship. So in my first model, I prefer to drop the variable *internation_stud.*

b) Regression Analysis for Model one

To determine the best predictive model for total score, I choose the model as below:

Total_score= β_0+ β_1(teaching) + β_2(international) + β_3(research) + β_4(citations) + β_5(income) + β_6 (num_students) + β_7(student_staff_ratio)+εi

Based on the output in table 2, Appendix, the model's P-value is very small and R^2 is very large, 0.9999. However, the variable num_students' P-value is pretty large (0.473). So I will do model selection in next step.

c) Model Selection

I use the stepwise selection in R in order to figure out a better fitted model. Based on the output in table 3 in appendix, I dropped the variable num_students. Then I will do the regression again by using the selected variables.

Model by stepwise:

Total_score= β_0+ β_1(teaching) + β_2(international) + β_3(research) + β_4(citations) + β_5(income) + β_6 (student_staff_ratio)+εi

Section3: Analysis of the Best Regression Equations

After analyzing the Pearson Correlation test and model selection, I dropped two variables named num_students and internation_stud. I used the regression analysis to figure out which one to be excluded from our equation. Below is the best regression equation for total score from those 6 selected :

total_score =-0.15 + 0.305(teaching)+ 0.075(international)+ 0.297(research)+ 0.301(citations)+ 0.024 (income) + 0.004(student_staff_ratio)

From the output in table 4, Appendix, it can be learned that R^2 is 0.9999. Since the multiple coefficient of determination (R squared) is close to 1, it presents a very good fit. The Variance Inflation Factors (VIF) values look reasonable since they are all under 10 (table 5, Appendix), which means there is no multicollinearity between the 6 variables.

In addition, I tested the assumption of multiple regression for the selected equation above. The first test of normal distribution for the error, showed by Figure 1a and b in Appendix, indicates the equation above is in fact a normal distribution with minimal outliers. The second assumption is to test the independence of the error. I used the Durbin-Watson (DW) Statistic to prove that there is no serial correlation and thus to prove the independence of the error. Using R we found the DW is equal to 2.201324 (table 6 in Appendix), which is close to 2 proving that the null hypothesis (ρϵ,ε-1=0) is reasonable. Therefore, the two assumptions are met.

Using the given Analysis of Variance (ANOVA) Table by R, we used the F test to evaluate the best regression equation for weight from those 6 selected variables. The hypothesis test is:

$$H_0: \beta_1=\beta_2=\beta_3=\beta_4=\beta_5=\beta_6=0 \textit{ VS } H_1: \text{at least one of the } \beta_i \text{ is not equal to zero.}$$

The F value is determined by the mean squared of regression divided by the mean squared of error. Since the computed value of F=2.204e+05 and the P-Value is < 2.2e-16 (Table 4), we reject the null hypothesis and conclude that the regression is significant at a level of 0.05. What's more, all the six variables are significant. Finally, we tested the Graphical Analysis of Residuals (residuals against fitted values) shown in Figure 1 c. We found that residual model against the fitted values is constant. This indicates the model is reasonable.

Section 4: Conclusion and Recommendation

From my analysis utilizing multiple methods of data processing technique, I have determined that the best acceptable models are applicable to the data. The best model includes factors such as teaching, international, research, citations, income, student_staff_ratio. And teaching, research, citations contribute much more than other factors to the ranking of a university. Based on the findings, if a university want to improve its' ranking, the management should pay more attention to teaching quality, research production and publication's citation.

5.Conclusion

a). United States has the most top universities and owns the top 1 university from 2011 to 2016.

b). The ranking of University of Cincinnati is on the list of top 200 universities in 2011 only.

c). The regression model developed here is reasonable. Based on this model, the most important factors affecting a university's ranking are teaching, research and citation.

6. Appendix:

Table 1

	teaching	international	research	citations	income	num_students	student_staff_ratio	international_students
teaching	1.00	-0.05	0.87	0.27	0.05	-0.09	-0.38	0.08
international	-0.05	1.00	0.09	0.11	-0.07	-0.24	0.16	0.82
research	0.87	0.09	1.00	0.21	0.16	0.01	-0.18	0.16
citations	0.27	0.11	0.21	1.00	-0.18	-0.23	-0.32	0.13
income	0.05	-0.07	0.16	-0.18	1.00	-0.01	0.02	-0.07
num_students	-0.09	-0.24	0.01	-0.23	-0.01	1.00	0.31	-0.32
student_staff_ratio	-0.38	0.16	-0.18	-0.32	0.02	0.31	1.00	-0.01
international_students	0.08	0.82	0.16	0.13	-0.07	-0.32	-0.01	1.00

n= 99

```
P
                       teaching international research citations income num_students student_staff_ratio international_students
teaching                               0.6535   0.0000    0.0073 0.6237       0.3838              0.0001                 0.4260
international            0.6535                 0.3971    0.2806 0.5166       0.0153              0.1182                 0.0000
research                 0.0000        0.3971             0.0383 0.1110       0.9303              0.0820                 0.1114
citations                0.0073        0.2806   0.0383           0.0680       0.0218              0.0013                 0.1838
income                   0.6237        0.5166   0.1110    0.0680              0.9106              0.8226                 0.5039
num_students             0.3838        0.0153   0.9303    0.0218 0.9106                           0.0019                 0.0011
student_staff_ratio      0.0001        0.1182   0.0820    0.0013 0.8226       0.0019                                     0.8925
international_students   0.4260        0.0000   0.1114    0.1838 0.5039       0.0011              0.8925

>
```

Table 2

```
Call:
lm(formula = total_score ~ teaching + international + research + 
    citations + income + num_students + student_staff_ratio, 
    data = d1)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.17288 -0.04856 -0.01263  0.03405  0.60981 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -1.228e-01  1.040e-01  -1.181    0.241    
teaching             3.045e-01  1.471e-03 207.038  < 2e-16 ***
international        7.480e-02  5.299e-04 141.157  < 2e-16 ***
research             2.970e-01  1.324e-03 224.404  < 2e-16 ***
citations            3.010e-01  9.081e-04 331.510  < 2e-16 ***
income               2.389e-02  4.050e-04  58.975  < 2e-16 ***
num_students        -6.095e-07  8.452e-07  -0.721    0.473    
student_staff_ratio  4.677e-03  1.074e-03   4.353 3.51e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.09122 on 91 degrees of freedom
Multiple R-squared:  0.9999,	Adjusted R-squared:  0.9999 
F-statistic: 1.879e+05 on 7 and 91 DF,  p-value: < 2.2e-16
```

Table 3

```
Stepwise Model Path 
Analysis of Deviance Table

Initial Model:
total_score ~ teaching + international + research + citations + 
    income + num_students + student_staff_ratio

Final Model:
total_score ~ teaching + international + research + citations + 
    income + student_staff_ratio


            Step Df    Deviance Resid. Df Resid. Dev       AIC
1                                      91  0.7572080 -466.4505
2 - num_students  1 0.004326731        92  0.7615347 -467.8864
```

Table 4

```
Call:
lm(formula = total_score ~ teaching + international + research + 
    citations + income + student_staff_ratio, data = d1)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.16422 -0.04674 -0.01225  0.03599  0.60756 
```

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.1501202  0.0966341  -1.553    0.124
teaching            0.3046112  0.0014524 209.724  < 2e-16 ***
international       0.0749299  0.0004986 150.273  < 2e-16 ***
research            0.2968649  0.0012964 228.985  < 2e-16 ***
citations           0.3011255  0.0008981 335.276  < 2e-16 ***
income              0.0239218  0.0004012  59.624  < 2e-16 ***
student_staff_ratio  0.0044759  0.0010350   4.324 3.87e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09098 on 92 degrees of freedom
Multiple R-squared:  0.9999,	Adjusted R-squared:  0.9999
F-statistic: 2.204e+05 on 6 and 92 DF,  p-value: < 2.2e-16
```

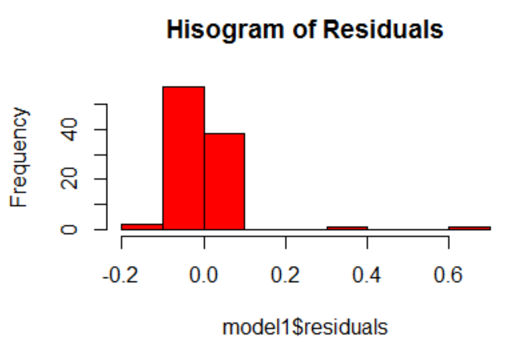
Table 5

```
teaching       international          research          citations            income      student_staff_ratio      5.542051          1.119002          5.120617          1.228735
1.125707          1.411789
```

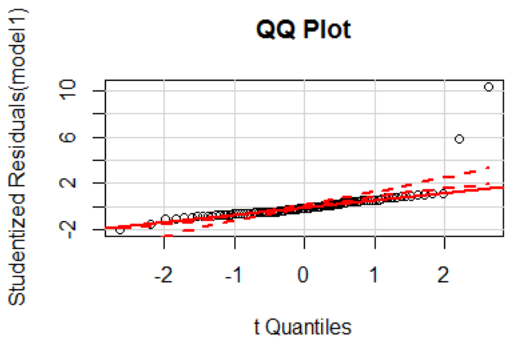
Table 6

```
> durbinWatsonTest(model1)
 lag Autocorrelation D-W Statistic p-value
   1      -0.1044311      2.201324   0.324
 Alternative hypothesis: rho != 0
```

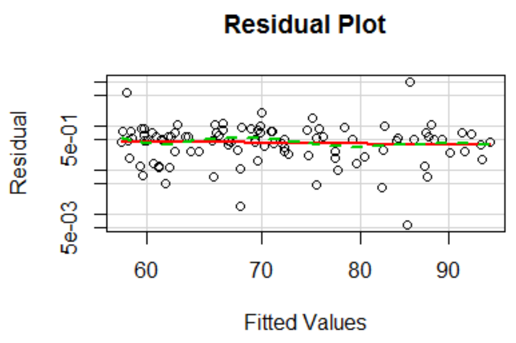
Fig 1



a



b



C

Code:

R Code

```
library(RSQLite)

library(RODBC)

odbcDataSources(type = c("all", "user", "system"))


#Create connection

db <- odbcConnect("Example", uid = "", pwd = "")

#Query a database (select statement)

UniversityRank <- sqlQuery(db, "SELECT * FROM WorldUniversityRanking.dbo.timesData")

sqlBasic <- sqlQuery(db, "SELECT
                world_rank
              ,university_name
              ,country
              ,teaching
              ,international
              ,research
              ,citations
              ,income
              ,total_score
              ,num_students
              ,student_staff_ratio
              ,international_students
              ,female_male_ratio
              ,year
              from WorldUniversityRanking.dbo.timesData")

library(dplyr)

# the summary of the obs
sqlsummary<-sqlQuery(db," Select Count(*) AS TotabObs, Avg(total_score) AS Avgscore
              From WorldUniversityRanking.dbo.timesData")
sqlsummary

#List the top 1 Universities from 2011 to 2016 */
TOP1<-sqlQuery(db,"select *
          from ( select ROW_NUMBER() over(partition by year order by world_rank ASC ) worldRank,
          university_name,country,year
          from WorldUniversityRanking.dbo.timesData
          where world_rank<201) a
          where worldRank<2")
TOP1
```

```
# rank of UC
UC<-sqlQuery(db," Select world_rank,university_name,total_score,year
            From WorldUniversityRanking.dbo.timesData
            where university_name like '%Cincinnati'")
UC

##Plot
library(ggplot2) # Data visualization
library(readr) # CSV file I/O, e.g. the read_csv function
library(dplyr)
library(plotly)
univer<-read.csv("C:/6045 R&SAS/final/data/timesData.csv",stringsAsFactors=FALSE,header=T);
str(univer)
#take Duke,Harvard,Princeton University as an example to show the change of rank
duke<-"Duke University"
duke.university<-univer[univer$university_name==duke,]

Harvard<-"Harvard University"
Harvard.University<-univer[univer$university_name==Harvard,]

Princeton<-"Princeton University"
Princeton.University<-univer[univer$university_name==Princeton,]

total <- rbind(duke.university,Harvard.University,Princeton.University)

str(total)
#converting world_rank

# converting world_rank in times to numeric

duke.university[,1]=as.numeric(duke.university$world_rank)
Harvard.University[,1]=as.numeric(Harvard.University$world_rank)
```

```
Princeton.University[,1]=as.numeric(Princeton.University$world_rank)
#https://plot.ly/r/line-and-scatter/
#Plotting rankings of the top 3 universities over the years
library(magrittr)
plot_ly(data= duke.university, x= ~year, y = ~world_rank,name ='duke', type='scatter', mode='lines+markers')%>%
  add_trace(data = Harvard.University,name= 'harvard')%>%
  add_trace(data=Princeton.University,name = 'princeton')%>%
  layout(title = 'University Ranking Over Time')

#read the data for the year of 2011
uni<-univer[univer$year==2011,]

# country VS university count
country_VS_uni<-uni %>%
  na.omit() %>%
  group_by(country)%>%
  summarize(count = n())

# List top 10 country
top_10_country <- country_VS_uni %>%
  arrange(desc(count)) %>%
  head(10)
top_10_country

# Plot top 10 country as per university count
ggplot(top_10_country,
       aes(x=reorder(country, -count), y=count, fill=country)) +
  geom_bar(stat="identity") +
  coord_flip() +
  theme(legend.position="none") +
  labs(x="Count",y="Country") +
  ggtitle("Top 10 Country by number of university ")
```

```
#mapping
library(rworldmap)
gtdMap <- joinCountryData2Map( country_VS_uni, nameJoinColumn="country", joinCode="NAME" )
mapParams <- mapCountryData(gtdMap,
                            nameColumnToPlot="count",
                            catMethod="fixedWidth",
                            numCats=4,colourPalette="rainbow",
                            mapTitle="The number of university VS Country")

#Regression
#due to the 2016 has the less missing value
uni_2016 <- univer[univer$year==2016,]
uni2016 <- uni_2016[1:100,4:12]
str(uni2016)
#output data
getwd()
library(RODBC)
write.csv(d, file = "C:/6045 R&SAS/final/data/tdclean2.csv", row.names = F, quote = F)
write.table(uni2016, file = 'C:/6045 R&SAS/final/data/tdclean.txt',quote = F)


#from the output, then it can be learned that some variable need to be cleaned.
#clean data
uni2016$international <- as.numeric(as.character(uni2016$international))
uni2016$income <- sub('-','0',uni2016$income)
uni2016$income <- as.numeric(as.character(uni2016$income))
uni2016$total_score <- as.numeric(as.character(uni2016$total_score))
uni2016$num_students <- gsub(',','',uni2016$num_students)
uni2016$num_students <- as.numeric(as.character(uni2016$num_students))
uni2016$international_students <- as.numeric(as.character(gsub('%','',uni2016$international_students)))/100
str(uni2016)
d1 <- na.omit(uni2016)
```

```
d1
summary(d1)

#Correlation
cor(d1[,c(1:5,7:9)],method="pearson")
#http://www.statmethods.net/stats/regression.html
#regression
model<- lm(total_score~., d1)
print(summary(model))

# Stepwise Regression
library(MASS)
model<- lm(total_score~teaching + international + research + citations +
        income + num_students +student_staff_ratio,d1)
print(summary(model))

#stepwise selection
step <- stepAIC(model, direction="both")
step$anova # display results


#Finnal Model
model1<-lm(total_score ~ teaching + international + research + citations +
        income + student_staff_ratio,d1)
print(summary(model1))

#Evaluate Collinearity
library(car)
vif(model1) # variance inflation factors

# Test for Autocorrelated Errors
library(lmtest)
```

```
durbinWatsonTest(model1)


#residual diagoues

#QQ plot

outlierTest(model1)# Bonferonni p-value for most extreme obs

qqPlot(model1, main="QQ Plot")

qqnorm(model1$residuals)


#Constant Variance

ncvTest(model1)

spreadLevelPlot(model1,main="Residual Plot",ylab="Residual")


plot(model1$fitted.values,model1$residuals,main="Fitted Value vs Residuals",xlab="Fitted Values",ylab="Residuals",col="red")


av.Plots(model1)

cutoff <- 4/((nrow(mtcars)-length(model1$coefficients)-2))

plot(model1, which=4, cook.levels=cutoff)


#diagouse

windows()

layout(matrix(c(1,2,3,4), 2, 2, byrow = TRUE))

qqnorm(model1$residuals)

plot(model1$fitted.values,model1$residuals,main="Fitted Value vs Residuals",xlab="Fitted Values",ylab="Residuals",col="red")

hist(model1$residuals,col="red",main = "Hisogram of Residuals")

plot(uni2016$total_score[1:99],model1$residuals,main="Observations vs Residuals",xlab="Observations",ylab="Residuals",col="blue")
```