R summer session final

Qiang Liu

目录

Personal information

南京大學

基于R语言的生物医学大数据分析报告

院　　系	生命科学学院
专　　业	生物科学
学　　号	191850112
年　　级	19級(大三)
学生姓名	刘强

Packages included

```
> library(dplyr)
> library(tinytex)
> library(openxlsx)
> library(ggplot2)
> library(psych)
> library(textreuse)
> library(skimr)
> library(tidyverse)
> library(knitr)
```

Evaluation of my performance

0)Create a list with 4 elements

```
> name <- c("Qiang Liu")
> gender <- c("Male")
> Mandatory1 <- c("FALSE","TRUE","TRUE","TRUE","TRUE","TRUE","TRUE",
+                 "FALSE","FALSE","FALSE","FALSE","FALSE")
> Score1 <- c(90,88,91,92,93,90,84,87,98,90,88,87)
> Course1 <- c("Sports Dance","Genetics","Experiments of Genetics","Physiology",
+              "Experiments of Physiology","Molecular Biology",
+              "Experiments of Molecular Biology",
+              "Polymeric Material and Social Development",
+              "History of the Peloponnesian War",
+              "The introduction to life science research hotspot",
+              "Analytical Biochemistry",
+              "Experiments of Analytical Biochemistry")
> Scores_of_last_semester <- data.frame(Mandatory1,Score1,row.names = Course1)
> colnames(Scores_of_last_semester) <- c("Mandatory","Score")
> Mandatory2 <- c("FALSE","TRUE","TRUE","TRUE","FALSE","FALSE","FALSE",
+                 "FALSE","FALSE","FALSE","FALSE","FALSE","FALSE")
> Score2 <- c(96,97,91,89,95,90,93,94,96,91,91,93,85)
> Course2 <- c("Immunology","Microbiology","Experiments of Microbology",
+              "Evolutional Biology","Evolution of Life
+              and Environmental Background",
+              "Strategies on Drug Discovery and Development",
+              "Pharmaceutics","Pharmaceutics Experiments",
+              "Specified English","Molecular Immunology",
+              "Drug Chemistry","Experiments of Drug Chemistry",
+              "Frontier in Pharmacological Sciences")
> Scores_of_this_semester <- data.frame(Mandatory2,Score2,row.names = Course2)
> colnames(Scores_of_this_semester) <- c("Mandatory","Score")
```

First of all, Name, Gender, and Courses name and grades of two semesters are defined as vectors. Then put all the vectors into the list.

The list is shown below.

```
> Qiang_Liu <- list(Name = name,Gender = gender,
+                   Scores_of_last_semester = Scores_of_last_semester,
+                   Scores_of_this_semester = Scores_of_this_semester)
> pander::pander(Qiang_Liu)
```

- **Name**: Qiang Liu
- **Gender**: Male
- **Scores_of_last_semester**:

	Mandatory	Score
Sports Dance	FALSE	90
Genetics	TRUE	88
Experiments of Genetics	TRUE	91
Physiology	TRUE	92
Experiments of Physiology	TRUE	93
Molecular Biology	TRUE	90
Experiments of Molecular Biology	TRUE	84
Polymeric Material and Social Development	FALSE	87
History of the Peloponnesian War	FALSE	98
The introduction to life science research hotspot	FALSE	90
Analytical Biochemistry	FALSE	88
Experiments of Analytical Biochemistry	FALSE	87

- **Scores_of_this_semester**:

	Mandatory	Score
Immunology	FALSE	96
Microbiology	TRUE	97
Experiments of Microbology	TRUE	91
Evolutional Biology	TRUE	89
Evolution of Life and Environmental Background	FALSE	95

	Mandatory	Score
Strategies on Drug Discovery and Development	FALSE	90
Pharmaceutics	FALSE	93
Pharmaceutics Experiments	FALSE	94
Specified English	FALSE	96
Molecular Immunology	FALSE	91
Drug Chemistry	FALSE	91
Experiments of Drug Chemistry	FALSE	93
Frontier in Pharmacological Sciences	FALSE	85

1) Have I got better scores this semester than last semester?

```
> Scorelast <- Qiang_Liu[[3]]
> Scorethis <- Qiang_Liu[[4]]
> pander::pander(summarise(Scorelast,mean_Score_of_last=mean(Score),
+                          sum_Score_of_last=sum(Score),
+                          sd_Score_of_last=sd(Score)))
```

mean_Score_of_last	sum_Score_of_last	sd_Score_of_last
89.83	1078	3.563

```
> pander::pander(summarise(Scorethis,mean_Score_of_this=mean(Score),
+                          sum_Score_of_this=sum(Score),
+                          sd_Score_of_this=sd(Score)))
```

mean_Score_of_this	sum_Score_of_this	sd_Score_of_this
92.38	1201	3.355

As can be seen from the tables, the average score of last semester is 89.83, while the average score of this semester is 92.38. And the standard deviation is smaller, which means that the difference between courses is smaller. Therefore, the grades of this semester have made great progress compared with last semester.

2) Which kind of course am I better at (mandatory vs optional)?

```
> Scoreall <- rbind(Scorelast,Scorethis)
> Scoreall_man <- filter(Scoreall,Mandatory==TRUE)
> Scoreall_opt <- filter(Scoreall,Mandatory==FALSE)
> pander::pander(summarise(Scoreall_man,mean_Score_of_Mandatory=mean(Score),
+                          sum_Score_of_Mandatory=sum(Score),
+                          sd_Score_of_Mandatory=sd(Score)))
```

mean_Score_of_Mandatory	sum_Score_of_Mandatory	sd_Score_of_Mandatory
90.56	815	3.575

```
> pander::pander(summarise(Scoreall_opt,mean_Score_of_Optional=mean(Score),
+                          sum_Score_of_Optional=sum(Score),
+                          sd_Score_of_Optional=sd(Score)))
```

mean_Score_of_Optional	sum_Score_of_Optional	sd_Score_of_Optional
91.5	1464	3.724

As can be seen from the tables, the average score of mandatory courses is 90.56 while that of optional courses is 91.5. Therefore, compared with the required courses, I got better grades in the optional courses. However, the standard deviation of optional courses is larger than mandatory courses, indicating that the differences between courses are greater.

---

3)What's the ranks of my first 10 courses this semester?

```
> My_data <- c('Me')
> My_data <- append(My_data,sort(Score2,decreasing=T)[1:10])
> My_data <- append(My_data,c('None','M'))
> All_Scores <- read.xlsx("Scores_this_semester.xlsx")
> All_Scores <- rbind(All_Scores,My_data)
> coursenum <- c(2:11)
> My_rank <- c('My rank')
> for (i in coursenum){
+   ranki <- All_Scores[order(All_Scores[,i],decreasing = TRUE),]
+   My_rank <- append(My_rank,which(ranki[,1]=='Me'))
+ }
> My_rank <- append(My_rank,c('None','M'))
> All_Scores <- rbind(All_Scores,My_rank)
> pander::pander(All_Scores[101:102,1:11])
```

表 7: Table continues below

	student	Course_1	Course_2	Course_3	Course_4	Course_5	Course_6	Course_7	Course_8	Course_9	Course_10
101	Me	97	96	96	95	94	93	93	91	91	91
102	My rank	6	5	11	10	21	20	19	37	31	22

The ranks of my first 10 courses this semester is shown in the table above.

Analyze students' performance

1)How many students got 10 As, 9 As, ... and 0 As (A is defined as >=90).

```
> Scores_of_students <- read.xlsx("Scores_this_semester.xlsx")
> pander::pander(table(rowSums(Scores_of_students[,2:11]>=90)))
```

0	1	2	3	4	5	6	7	8
2	5	9	24	27	16	11	5	1

As can be seen from the table, 1 student got 8As, 5 students got 7As, 11 students got 6As, 16 students got 5As, 27 students got 4As, 24 students got 3As, 9 students got 2As and 5 students got 1A. Only 2 students didn't get any A.

2)“Birds of a feather flock together”.

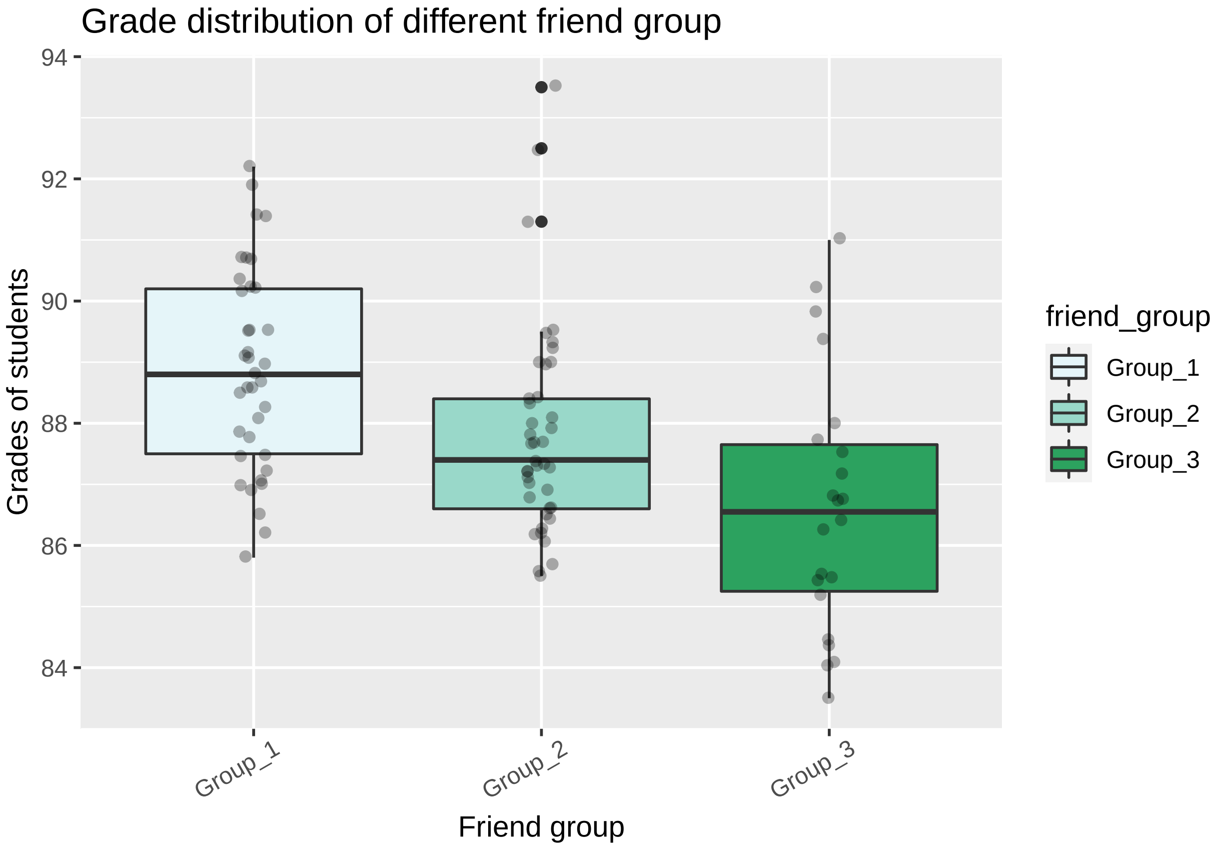
```
> Average<-apply(Scores_of_students[c(2:11)],1,mean)
> Scores_of_students <- data.frame(Scores_of_students,Average)
> pander::pander(aggregate(Scores_of_students$Average,
+           by=list(friendgroup=Scores_of_students$friend_group),mean))
```

friendgroup	x
1	88.89
2	87.83
3	86.63

```
> m <- c(1:100)
> a <- c("Group_1")
> b <- c("Group_2")
> c <- c("Group_3")
> for (i in m){
+   if (Scores_of_students[i,12]==1){
+     Scores_of_students[i,12] <- a
+   }
+   if (Scores_of_students[i,12]==2){
+     Scores_of_students[i,12] <- b
+   }
+   if (Scores_of_students[i,12]==3){
+     Scores_of_students[i,12] <- c
+   }
+ }
```

As can be seen from the table, friend group 1 got the best average grade and friend group 3 got the lowest average grade.

```r
> ggplot(Scores_of_students,aes(x=friend_group,y=Average)) +
+   geom_boxplot(aes(fill=friend_group)) +
+   geom_jitter(width = 0.05, alpha = 0.3, color = 'black') +
+   labs(title="Grade distribution of different friend group",
+        x="Friend group",
+        y = "Grades of students") +
+   theme(axis.text.x = element_text(angle = 30,vjust = 0.85,hjust = 0.75))+
+   scale_fill_brewer(type="seq",palette = 2)
```



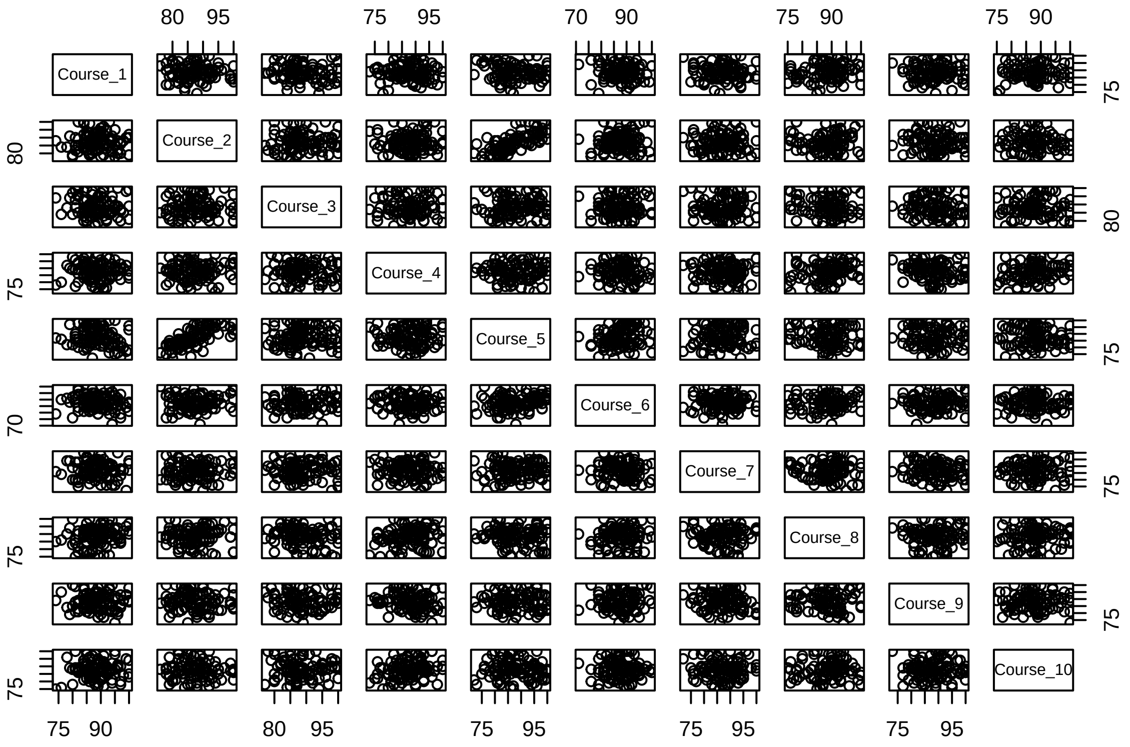
As can be seen from the graph, overall, friend group 1 had the highest grade distribution and friend group 3 had the lowest grade distribution. But in terms of individual performance, there were also high performers in each group. For example, Group 2 had the highest average score. So I’m partly agree with the saying “Birds of a feather flock together” , in general it’s right but not in all conditions.

---

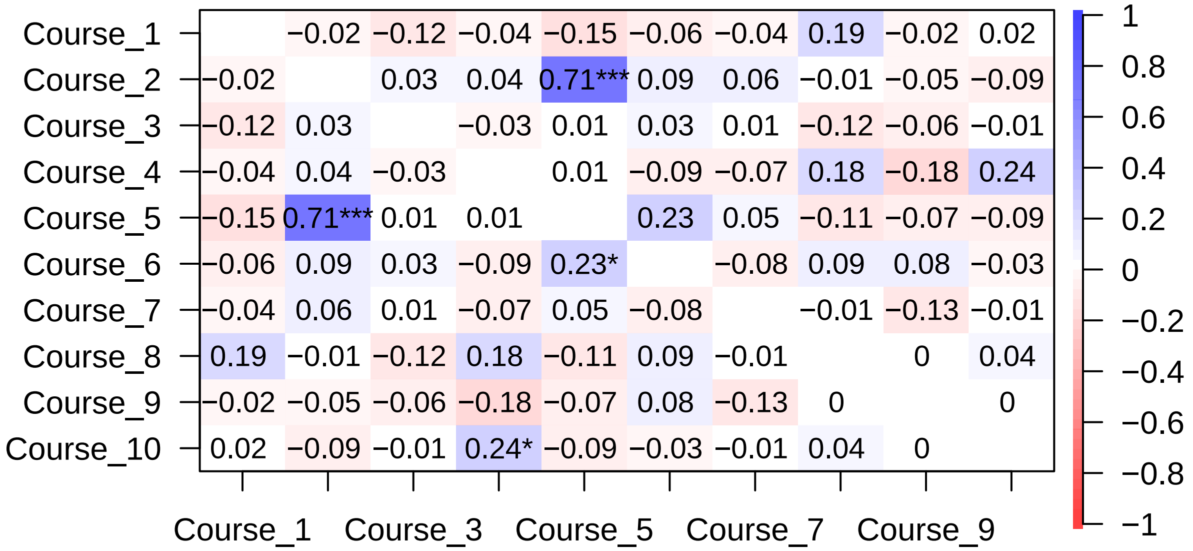
3)Any two or more courses might be highly related to each other?

First see the rough correlations between all Courses shown below.

```
> correlation <- corr.test(Scores_of_students[,2:11])
> r_value <- correlation$r
> p_value <- correlation$p
> pairs(Scores_of_students[,2:11])
```



```
> corPlot(r_value,pval=p_value,numbers=TRUE,diag=FALSE,stars=TRUE)
```



As can be seen from the graph, all correlations between courses are shown. And the bond between Course2 and Course5 is robust.

First fit the linear regression model with two courses, and use ggplot to draw the linear regression.

```
> full.model <- lm(Course_2 ~ Course_5,data = Scores_of_students)
> pander::pander(summary(full.model))
```

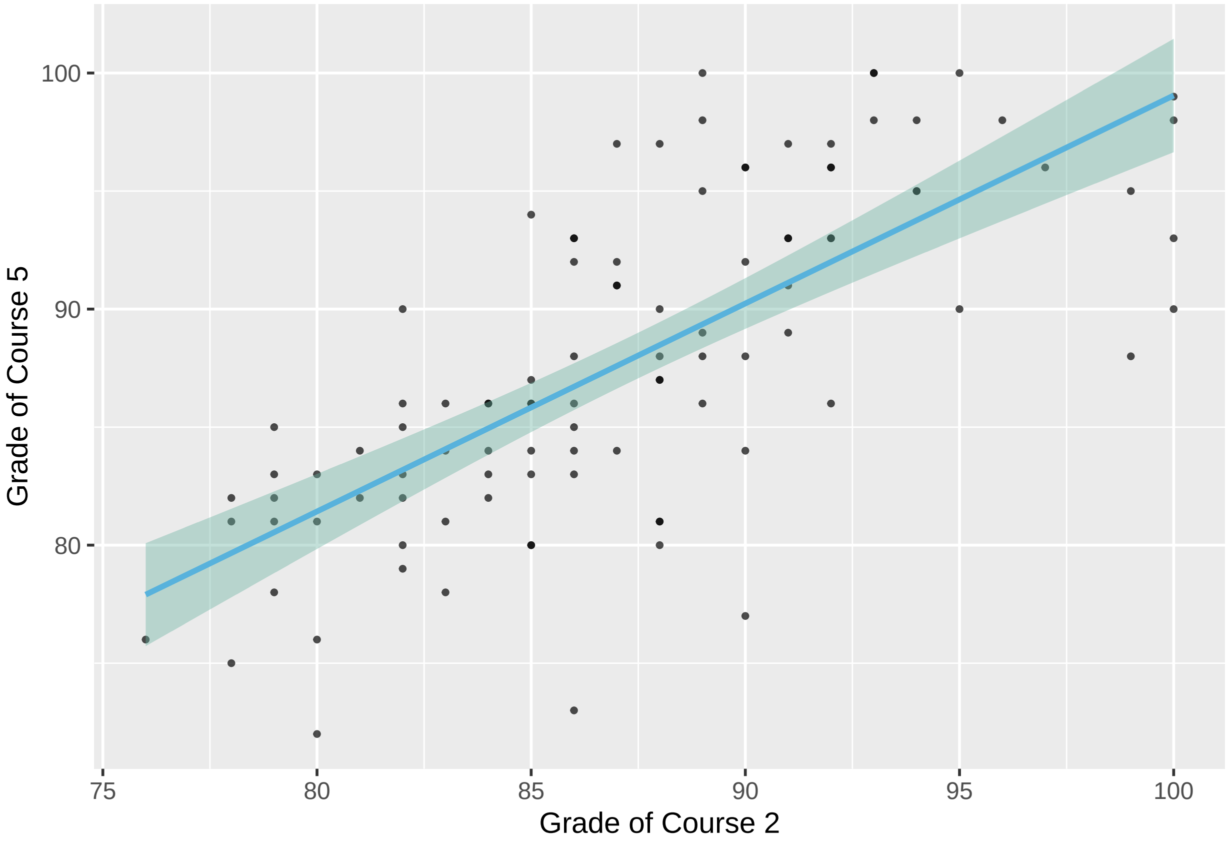
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.43	5.051	7.213	1.162e-10
Course_5	0.579	0.05731	10.1	7.298e-17

表 12: Fitting linear model: Course_2 ~ Course_5

Observations	Residual Std. Error	R^2	Adjusted R^2
100	3.945	0.5101	0.5051

```
> Scores_of_students%>%
+   ggplot(aes(x = Course_2,y = Course_5))+
+   geom_point(colour = "black",size = 0.7,alpha = 0.7)+
+   geom_smooth(method = lm, color = "#58B2DC",fill = "#69b3a2",se = TRUE)+
+   xlab("Grade of Course 2")+
+   ylab("Grade of Course 5")+
+   scale_fill_brewer(type="seq",palette = 2)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



And we can get the interpretation of these two courses as shown below.

$$\hat{Course5} = 36.43 + 0.579 * Course2$$

******

4)During exams of Course 1-10, someone sit side-by-side with Student_27 and copied quite a few answers from Student_27, any clue who he is?

```
> row <- c(1:100)
> row <- row[-27]
> col <- c(2:11)
> for (i in row){
+   x = 0
+   for (j in col){
+     if(Scores_of_students[27,j]==Scores_of_students[i,j]){
+       x = x+1
+     }
+   }
+   if (x>=2){
+     dex <- c(i,x)
+     print(dex)
+   }
+ }
```

```
## [1] 25  2
## [1] 34  3
## [1] 49  2
## [1] 51  2
## [1] 53  2
## [1] 58  2
## [1] 79  2
```

I tried to find the cheater by searching for people who got the same grade of same course with student 27, and I think the one who have more counts will be more likely to be judged as cheater. As is shown in above, I think students 34 might be the one who sit side-by-side with student 27 and copied quite a few answers from student 27, because he got 3 courses with the same grade with student 27.

---

5)Description statistics of each course and which course is much more difficult?

```
> pander::pander(summary(Scores_of_students[2:11]))
```

表 13: Table continues below

Course_1	Course_2	Course_3	Course_4
Min. : 74.00	Min. : 76.00	Min. : 77.00	Min. : 73.00
1st Qu.: 85.00	1st Qu.: 83.00	1st Qu.: 85.00	1st Qu.: 83.75
Median : 88.00	Median : 87.00	Median : 88.00	Median : 88.00
Mean : 88.54	Mean : 87.29	Mean : 88.68	Mean : 87.68
3rd Qu.: 92.00	3rd Qu.: 91.00	3rd Qu.: 92.25	3rd Qu.: 92.00
Max. :100.00	Max. :100.00	Max. :100.00	Max. :100.00

表 14: Table continues below

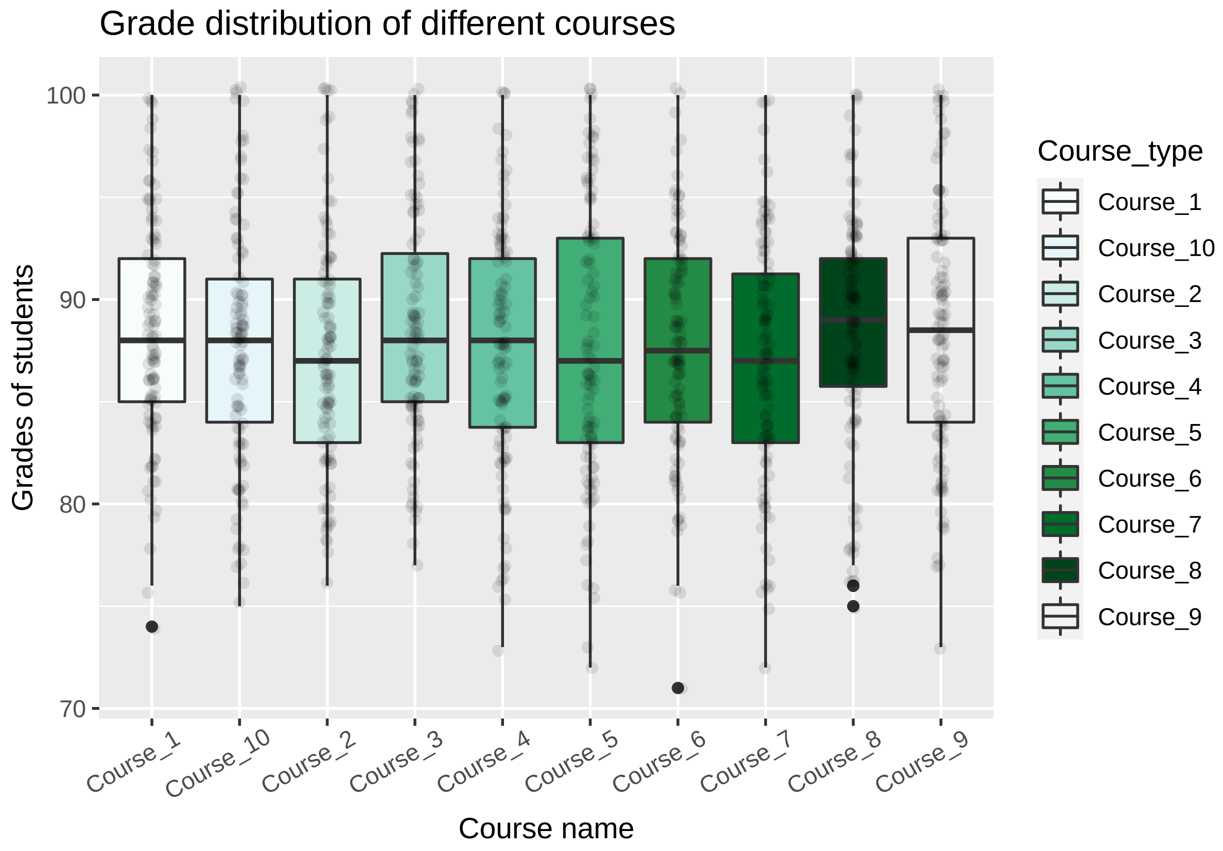
Course_5	Course_6	Course_7	Course_8
Min. : 72.00	Min. : 71.00	Min. : 72.00	Min. : 75.00
1st Qu.: 83.00	1st Qu.: 84.00	1st Qu.: 83.00	1st Qu.: 85.75
Median : 87.00	Median : 87.50	Median : 87.00	Median : 89.00
Mean : 87.85	Mean : 87.67	Mean : 87.10	Mean : 88.53
3rd Qu.: 93.00	3rd Qu.: 92.00	3rd Qu.: 91.25	3rd Qu.: 92.00
Max. :100.00	Max. :100.00	Max. :100.00	Max. :100.00

Course_9	Course_10
Min. : 73.00	Min. : 75.00
1st Qu.: 84.00	1st Qu.: 84.00
Median : 88.50	Median : 88.00
Mean : 88.42	Mean : 87.82
3rd Qu.: 93.00	3rd Qu.: 91.00
Max. :100.00	Max. :100.00

The description statistics(Min, 1st Quater, Median, Mean, 3rd Quater and Max) of each course are shown above in the tables.

```
> Scores_of_students.new <- Scores_of_students%>%
+   pivot_longer(c(Course_1,Course_2,Course_3,Course_4,Course_5,
+                  Course_6,Course_7,Course_8,Course_9,Course_10),
+                names_to = "Course_type", values_to = "Grade")
> ggplot(Scores_of_students.new,aes(x=Course_type,y=Grade)) +
+   geom_boxplot(aes(fill=Course_type)) +
+   geom_jitter(width = 0.05, alpha = 0.1, color = 'black') +
+   labs(title="Grade distribution of different courses",
+        x="Course name",
+        y = "Grades of students") +
+   theme(axis.text.x = element_text(angle = 30,vjust = 0.85,hjust = 0.75)) +
+   scale_fill_brewer(type="seq",palette = 2)
```

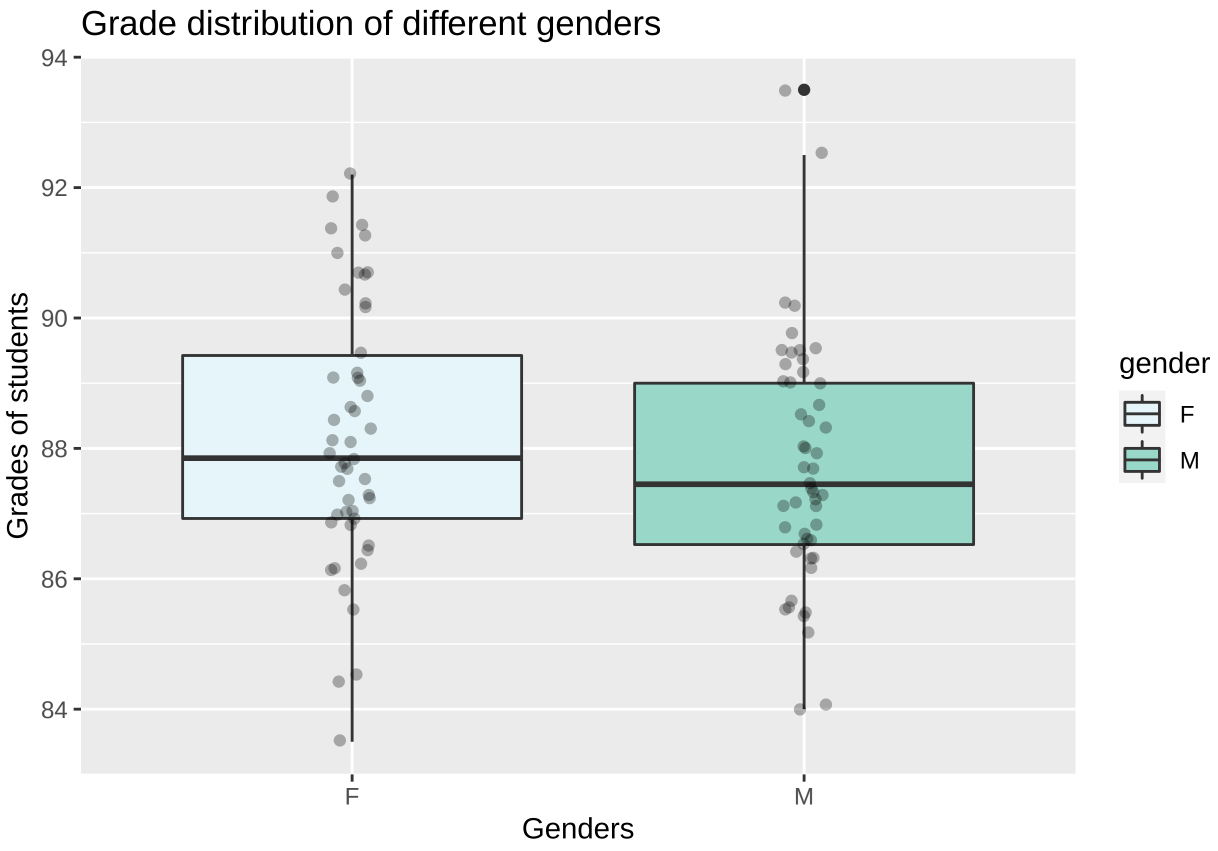
```
## Warning in RColorBrewer::brewer.pal(n, pal): n too large, allowed maximum for palette BuGn is 9
## Returning the palette you asked for with that many colors
```



As can be seen from the boxplot, we can see the grade distribution of each course. In general, the means of all courses are about 87 88. In all 10 courses, course 7 got the lowest average grade of 87.10, so maybe course 7 is much more diffcult.

6)Is performance different between different genders?

```
> ggplot(Scores_of_students,aes(x=gender,y=Average)) +
+   geom_boxplot(aes(fill=gender)) +
+   geom_jitter(width = 0.05, alpha = 0.3, color = 'black') +
+   labs(title="Grade distribution of different genders",
+        x="Genders",
+        y = "Grades of students") +
+   scale_fill_brewer(type="seq",palette = 2)
```



As can be seen from the boxplot, I used average to estimate the performance of each students. Generally the distribution of Female students are higher than Male students, and you can see more points of high average in the Female group. But in terms of individual performance, there were also high performers in Male group. So I agree that there's a difference in performance between different genders.

课程建议

因为之前学过生物统计学这门课程，对 R 熟悉程度高一点，一些包的调用有点印象，所以初期学起来还是比较轻松的，不过后面开始讲比较深的生信相关的代码的时候跟着有点吃力，可能没有相关专业背景的话听起来会比较难懂。税昆明学长讲的单细胞测序部分挺好的感觉，代码讲的比较慢。其实没啥特别的建议，老师教的很好，可能设备原因是我碰到的比较大的问题。一是感觉投屏看不清然后可能感觉部分代码可以用截图的方式用 ppt 讲，因为 Rstudio 页面和字体还是有点小。二是跟着跑代码像我的电脑一会就没电了，可能和之前群里讨论的一样开设在有充电口的教室会比较方便。老师的课很棒，短时间输入这么多知识本来就很难，希望之后的课程能够顺利开展。