

14조_3차발표

2022-06-04

< 프랑스 이민자 노동시장과 인종차별 >

0. 패키지 로딩

```
# 패키지 설치
#install.packages("ggplot2")
#install.packages("readr")
#install.packages("RColorBrewer")
#install.packages("data.table")
#install.packages("gmodels")
#install.packages("blme")
#install.packages("effects")
```

```
library(blme)
```

```
## 필요한 패키지를 로딩중입니다: lme4
```

```
## 필요한 패키지를 로딩중입니다: Matrix
```

```
library(effects)
```

```
## 필요한 패키지를 로딩중입니다: carData
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
library(gmodels)
library(ggplot2)
library(readr)
library(RColorBrewer)
library(data.table)
```

1. 2차발표 피드백 반영 및 수정사항

2. 2018년 이민자와 비이민자의 비율 비교-신지원

2.1. 데이터 불러오기

```
#새로운 데이터 프레임 형성
immi_total <- read.csv("Donnees_complementaires_total.csv")
```

2.2.데이터 전처리

```

immi_total <- immi_total[,c(1,2,4,6)] #전체 인구 수, 이민자 수 추출
immi_total <- immi_total[-c(1,2,3),]

#열이름을 통해 "연도", "외국인수", "이민자수", "전체인구" 설정
colnames(immi_total) <- c("연도", "외국인수", "이민자수", "전체인구")

# 문자형을 숫자형으로 변환
for(i in 2:4){
  immi_total[,i] <- gsub(",", "", immi_total[,i])
  immi_total[,i] <- as.numeric(immi_total[,i])
}

immi_total$연도 <- gsub("(p)", "", immi_total$연도) # 연도의 추정치 표시 제거

```

2.3. 2018년 이민자의 비율 데이터 시각화

```

#2018 이민자수와 비이민자수 비교
comp_immi <- subset(immi_total, 연도==2018)

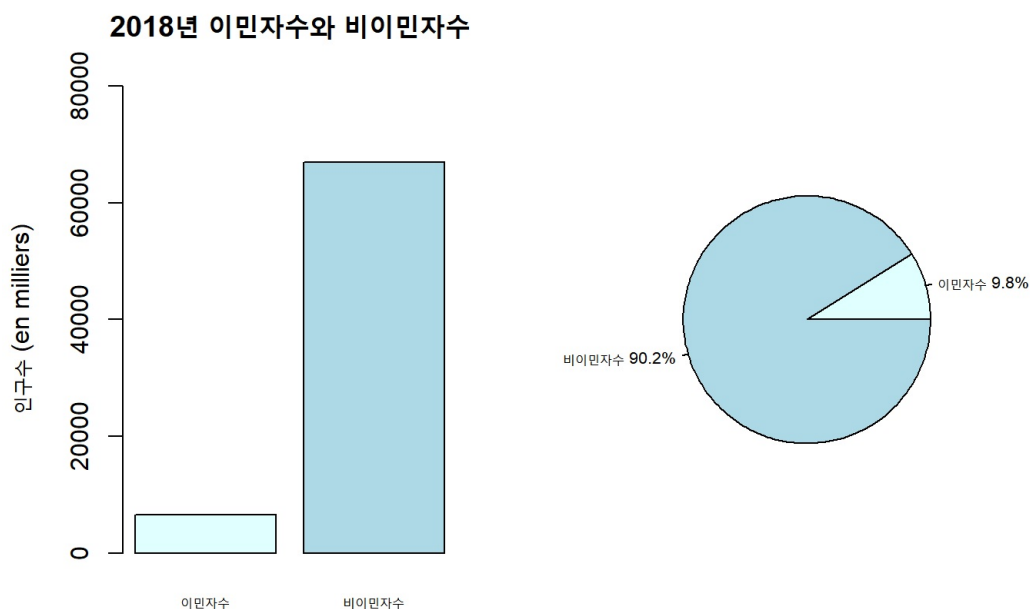
unlist_df <- comp_immi[,c(3,4)]
unlist_df <- unlist(unlist_df)

opar = par(mfrow=c(1,2))
barplot(unlist_df, names=c("이민자수", "비이민자수"), ylim=c(0,80000), cex.names=0.65,
        col = c("lightcyan", "lightblue"))

title(main="2018년 이민자수와 비이민자수", ylab="인구수 (en milliers)")
lab <- c("이민자수", "비이민자수")
lab <- paste(lab, c(round(comp_immi$이민자수*100.00/comp_immi$전체인구, 1),
                    round(100 - (comp_immi$이민자수*100.00/comp_immi$전체인구), 1)))
lab <- paste(lab, "%", sep="")

pie(unlist_df, labels=lab, main="", col = c("lightcyan", "lightblue"), cex = 0.65)

```



3. 출신대륙별 이민자 수 추이

3.1. 2018년 출신대륙별 이민자의 비율-신지원

(1) 2018년 데이터 불러오기

- 출처: <https://www.insee.fr/fr/statistiques/5366472?sommaire=5363676&q=immigrant+2018> (<https://www.insee.fr/fr/statistiques/5366472?sommaire=5363676&q=immigrant+2018>)

```
#새로운 데이터 프레임 형성
library(RColorBrewer)
df1 <- read.csv("pays_naissance_detailed_1968_2018.csv")

library(readr)
#df1 <- read_delim("pays_naissance_detailed_1968_2018.csv",
#delim = "\t", escape_double = FALSE,
#trim_ws = TRUE)
```

(2) 데이터 전처리

```
df1 <- df1[,1:2]
df1 <- df1[-c(1,2),]

#열이름을 통해 "pays", "population"로 설정
colnames(df1) <- c("pays", "population")
df1$population <- gsub(",","", df1$population)
df1$population <- as.integer(df1$population)

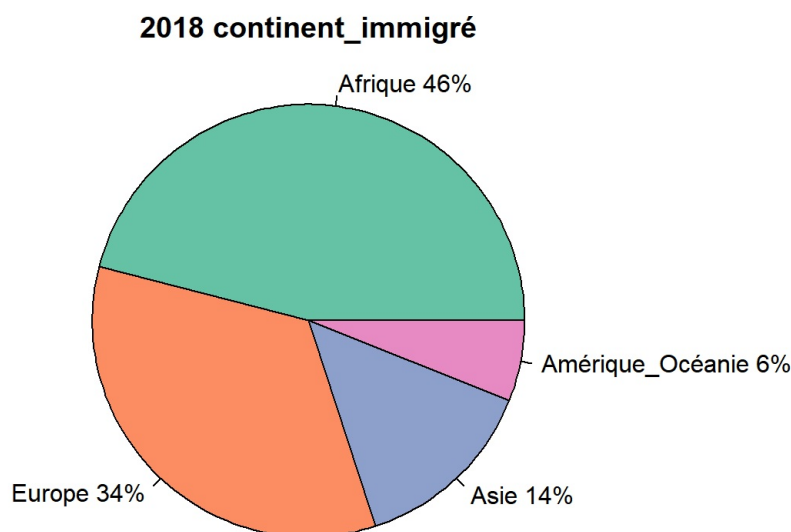
#"continent_population"라는 데이터 프레임 형성
continent_population <- df1[c(1,53,96,143),]
colnames(continent_population) <- c("pays","population")
head(df1)
```

```
##      pays population
## 3 Afrique    3034000
## 4 Algérie    838800
## 5 Maroc     784900
## 6 Tunisie    290900
## 7 Comores    125200
## 8 Sénégal    110500
```

```
#character에서 integer로 변환
continent_population_proportion <- round(100*prop.table(continent_population$population))
```

(3) 데이터 시각화

```
#대륙별 파이차트 만들기(lbl을 사용해 새롭게 벡터 형태로 저장. paste0를 사용하여 "lbl"과 "대륙별 인구수 비율"을 합쳐줌.)
lbl <- c("Afrique","Europe","Asie","Amérique_Océanie")
lbl_continent_population_proportion <- paste(lbl, continent_population_proportion)
lbl_continent_population_proportion <- paste0(lbl_continent_population_proportion,"%")
pie(continent_population_proportion, labels =lbl_continent_population_proportion, col=brewer.pal(4,"Set2"), main="2018 continent_immigré", init.angle=0, radius=1.0)
```



3.2. 2019년 출신대륙별 이민자 데이터 - 지명은

(1) 데이터 불러오기

```
immigrant2019 <- read.csv("demo-etran-origine-geo-immig2019.csv", encoding = "UTF-8")
```

(2) 데이터 전처리

데이터의 열이름 정리 및 필요없는 행과 열 삭제하여 아프리카, 유럽, 아시아, 아메리카, 오세아니아 다섯 대륙의 정보만 추출하였다.

```
names(immigrant2019) <- c(immigrant2019[2,1],immigrant2019[3,2],immigrant2019[3,3])
immigrant2019 <- immigrant2019[c(-1,-2,-3),]

#
# 대륙별 정보만 추출
immi19 <- immigrant2019[, c(1:3)]
immi19 <- immigrant2019[c(1, 13, 25, 37), c(1:3) ]

names(immi19) <- c("continent", "immigrant", "proportions")
```

- immi19 데이터 프레임 확인

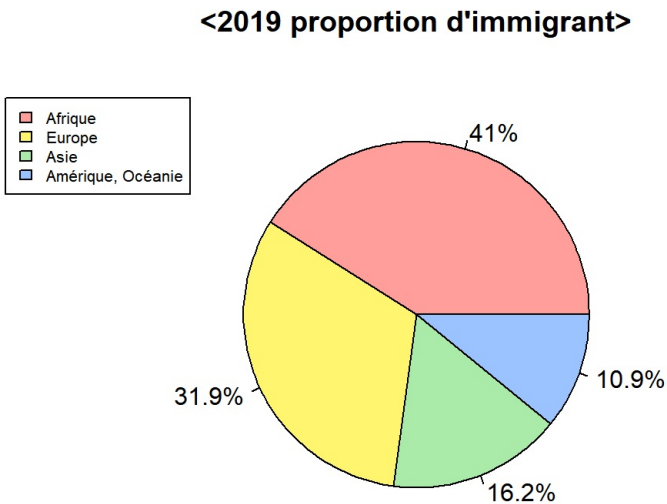
	continent	immigrant	proportions
4	Afrique	111,700	41
16	Europe	87,000	31.9
28	Asie	44,000	16.2
40	Amérique, Océanie	29,700	10.9

(3) 2019년 이민자 데이터 시각화하기

pie chart를 이용하여 2019년 출신대륙별 이민자의 비율 시각화하였다. 2019년도의 경우 아프리카 대륙 출신의 이민자(41%)가 가장 많았으며, 아프리카 대륙다음으로는 유럽대륙(31.9%)이 그 다음으로 많았다.

```
col1 <-c("#FF9E9B", "#FFF56E", "#AAEBAA", "#9BC3FF")
immi19$proportions<- as.numeric(immi19$proportions) # 숫자형 변환
i19_labels <- paste(immi19$proportions, "%", sep="") # 퍼센트 추가

pie(immi19$proportions, labels=i19_labels, col=col1,
    main="<2019 proportion d'immigrant>" #파이차트 생성
    legend(-1.9,1,"topleft", legend=immi19$continent, # 범례추가
    fill = col1, cex= 0.65)
```



3.3. 2020년 출신대륙별 이민자 데이터-지명은

(1) 데이터 불러오기

- 데이터 출처: <https://www.insee.fr/fr/statistiques/3633212#graphique-infographie> (<https://www.insee.fr/fr/statistiques/3633212#graphique->

infographie) 에서 발췌한 csv 파일에서 Fugure3을 사용하였다. 참고) 2020년 이민자의 수는 215,200명

```
immigrant2020 <- read.csv("Donnees_complementaires2020.csv", encoding = "UTF-8", header = T)
```

(2) 데이터 전처리

위 과정은 2019년 데이터 정제방법과 동일하다

```
names(immigrant2020) <- c(immigrant2020[2,1],immigrant2020[3,2],immigrant2020[3,3])
immigrant2020 <- immigrant2020[c(-1,-2,-3),c(1,2,3)]
```

```
immi20 <- immigrant2020[, c(1:3)]
immi20 <- immigrant2020[c(1, 13, 25, 37), c(1:3) ]
names(immi20) <- c("continent", "immigrant", "proportions")
```

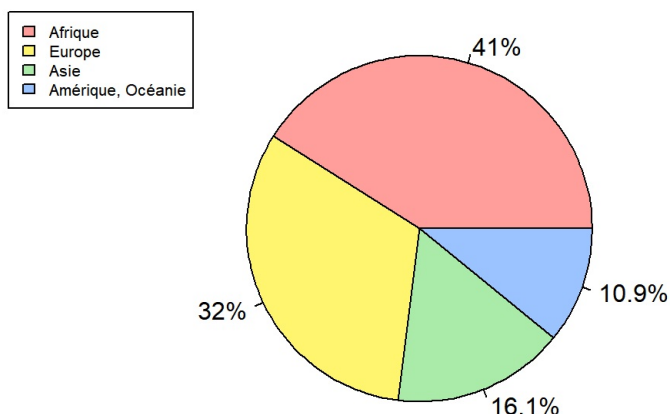
(3) 2020년도 이민자 데이터 시각화하기

pie chart를 이용하여 시각화를 진행하였다. 그 결과 앞서 2019년과 마찬가지로 2020년도에도 아프리카 대륙 출신 이민자가 가장 많았으며 유럽대륙 출신의 이민자가 그 뒤를 이었다.

```
immi20$proportions<- as.numeric(immi20$proportions) # 숫자형 변환
i20_labels <- paste(immi20$proportions, "%", sep="") # 퍼센트 추가

pie(immi20$proportions, labels=i20_labels, col=col1, main="<2020 proportion d'immigrant>") #파이차트 생성
legend(-1.9,1,"topleft", legend=immi20$continent, # 범례추가
      fill = col1, cex= 0.65)
```

<2020 proportion d'immigrant>



3.4. 2021년 출신대륙별 이민자 데이터-지명은

(1) 2021년 데이터 불러오기

- 데이터 출처: <https://www.insee.fr/fr/statistiques/3633212#graphique-infographie> (https://www.insee.fr/fr/statistiques/3633212#graphique-infographie) 에서 발췌한 csv 파일에서 Fugure2를 사용하였다.

참고) 2021년 이민자의 수는 6964(en milliers)명

(2) 데이터 전처리

```
# 데이터의 열이름 정리 및 필요없는 행열 삭제
immi21 <- c(immi21[c(3,55,98,145),c(1,2,3)])
names(immi21) <- c("continent","2018","2017")

immi21<-as.data.frame(immi21)

for(i in 2:3){
  immi21[,i] <-gsub("","",immi21[,i]) # 콤마 삭제
  immi21[,i] <-as.numeric(immi21[,i]) # 숫자형으로 변환
}

immi21$immigrant <- immi21$X2018 - immi21$X2017
# 2018년도 총 이민자에서 2017년도 총 이민자수를 2018년도에 유입된 이민자 인구를 추출

immi21 <- immi21[,-c(2,3)]

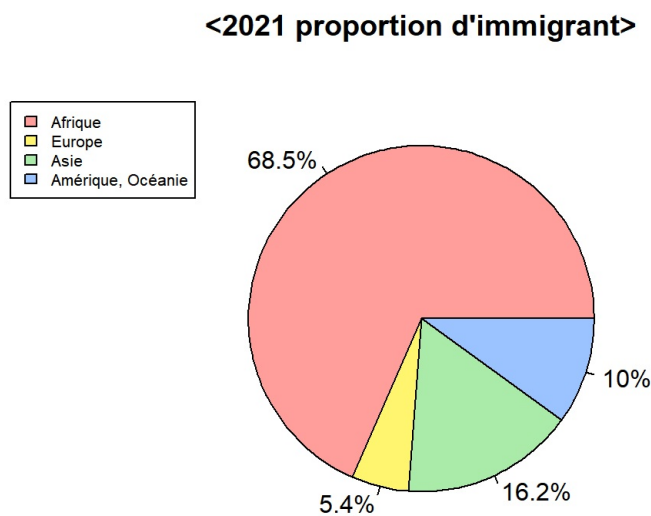
immi21$proportions <- c(round(immi21$immigrant[1]/sum(immi21$immigrant)*100,1), # 비율 데이터 추가
  round(immi21$immigrant[2]/sum(immi21$immigrant)*100,1),
  round(immi21$immigrant[3]/sum(immi21$immigrant)*100,1),
  round(immi21$immigrant[4]/sum(immi21$immigrant)*100,1))
```

(3) 2021년 이민자 데이터 시각화하기

마지막으로 2021년에는 아프리카 대륙 출신 이민자가 가장 많았으며 그 다음으로는 유럽 대륙의 이민자가 가장 많은 것으로 확인되었다. 2021년에는 다른 연도보다 아프리카 출신의 이민자 비율이 상승했음을 확인할 수 있었다.

```
i21_labels <- paste(immi21$proportions, "%", sep="") # 퍼센트 추가
pie(immi21$proportions, labels=i21_labels,col=coll,
  main="2021 proportion d'immigrant") #파이차트 생성

legend(-1.9,1,"topleft", legend=immi21$continent, # 범례추가
  fill = coll, cex= 0.65)
```



3.5. 대륙별 2019-2021년 이민자 수 추이-지명은

마지막으로 대륙별 전체적인 추이를 한 눈에 알아보도록 하겠다 먼저 그래프를 그리기 위한 전처리 과정이다

(1) 데이터 전처리

2019년부터 2021년의 데이터를 하나의 data frame으로 합친 후 세부 대륙별로 각각 데이터 프레임을 생성하였다.

```

immi19_21 <- rbind(immi19, immi20, immi21)
immi19_21$annee <- c("2019","2019","2019","2019",
                     "2020","2020","2020","2020",
                     "2021","2021","2021","2021")
immi19_21$immigrant <- gsub(" ", "", immi19_21$immigrant)
immi19_21$immigrant <- as.numeric(immi19_21$immigrant)
immi19_21$annee <- as.numeric(immi19_21$annee)
immi19_21$continent[c(4,8,12)] <- "Amérique, Océanie" # 세부 데이터 수정

```

```

afri_immi <- immi19_21[c(1,5,9),] # 아프리카 대륙
Eur_immi <- immi19_21[c(2,6,10),] # 유럽 대륙
Asie_immi<- immi19_21[c(3,7,11),] # 아시아 대륙
Amer_immi <- immi19_21[c(4,8,12),] # 아메리카 오세아니아 대륙

```

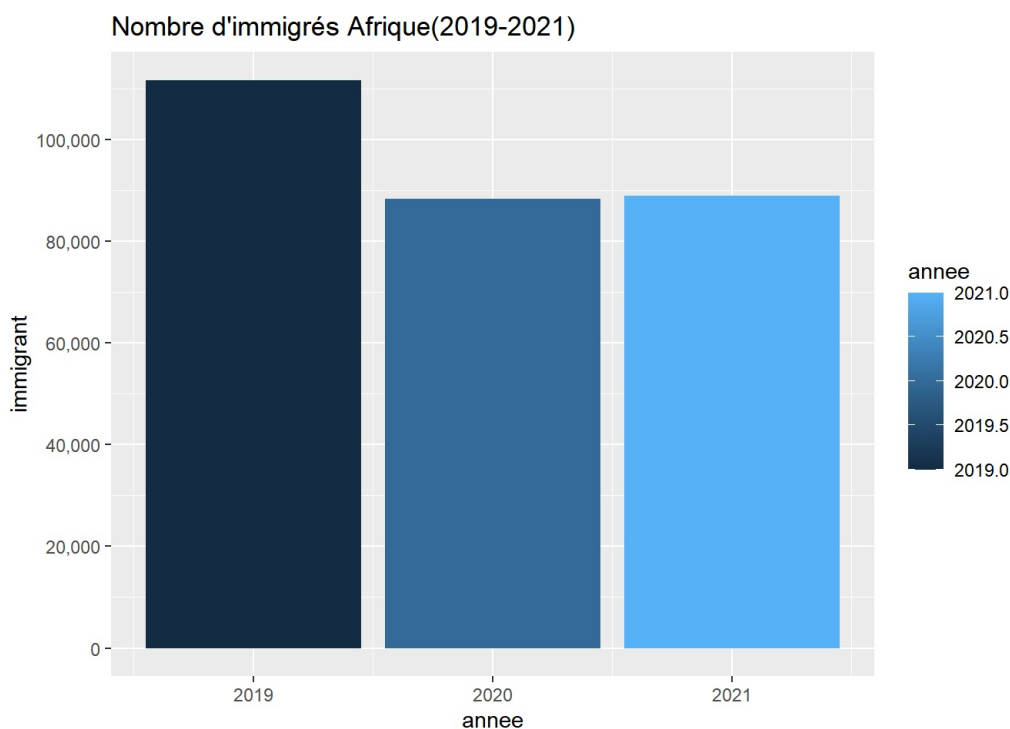
(2) 아프리카 연도별 추이

그래프를 통해 확인해본 결과 3개년 중 19년도의 아프리카 대륙출신 이민자가 가장 많았으며 2021년에 전년대비 소폭 상승한것을 확인할 수 있다.

```

ggplot(afri_immi, aes(x=annee, y=immigrant, fill=annee))+
  geom_col() +
  labs(title="Nombre d'immigrés Afrique(2019-2021)")+
  scale_y_continuous(labels = scales::comma, breaks = seq(0,100000,20000))

```



(3) 유럽 대륙의 추이

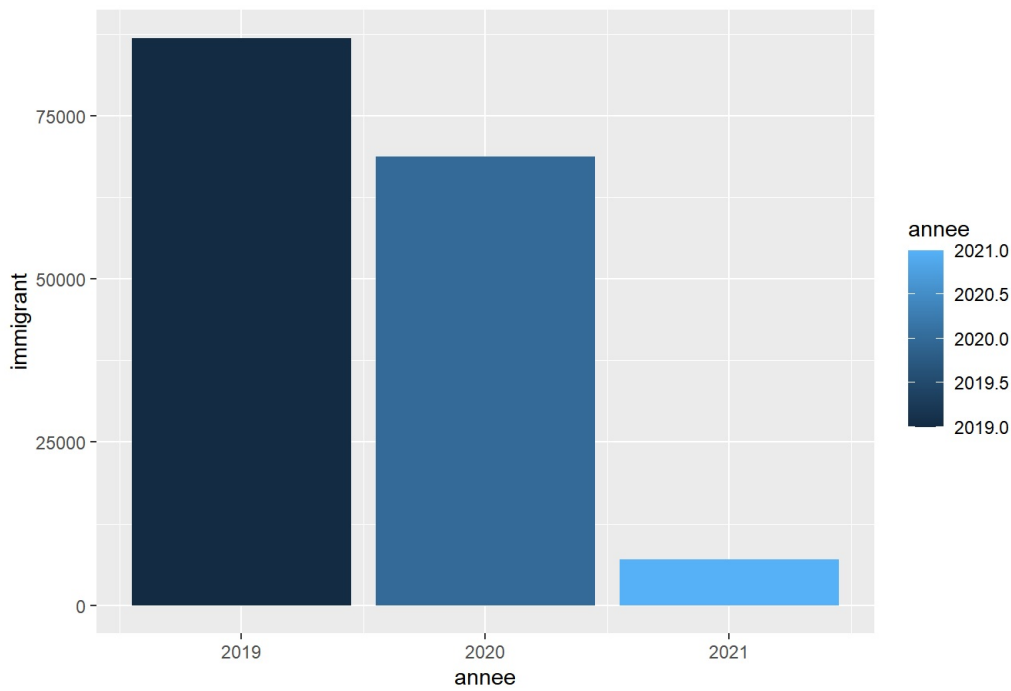
다음으로는 유럽 대륙 출신 이민자 수를 그래프를 통해 시각화해 보았다. 아프리카 대륙과 마찬가지로 19년도에 가장 많았던 이민자수는 21년 가파르게 하락하는 양상을 보였다.

```

ggplot(Eur_immi, aes(x=annee, y=immigrant, fill=annee))+
  geom_col() +
  labs(title="Nombre d'immigrés Europe(2019-2021)")

```

Nombre d'immigrés Europe(2019-2021)

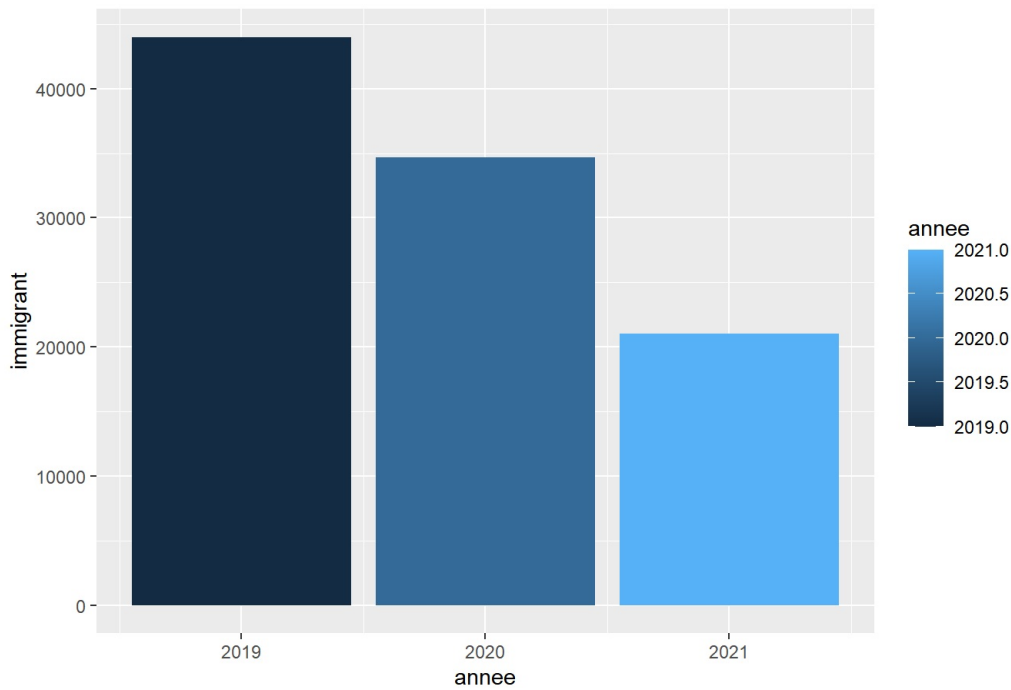


(4) 아시아 연도별 추이

아시아 대륙 또한 점차 하락하는 양상을 보였다.

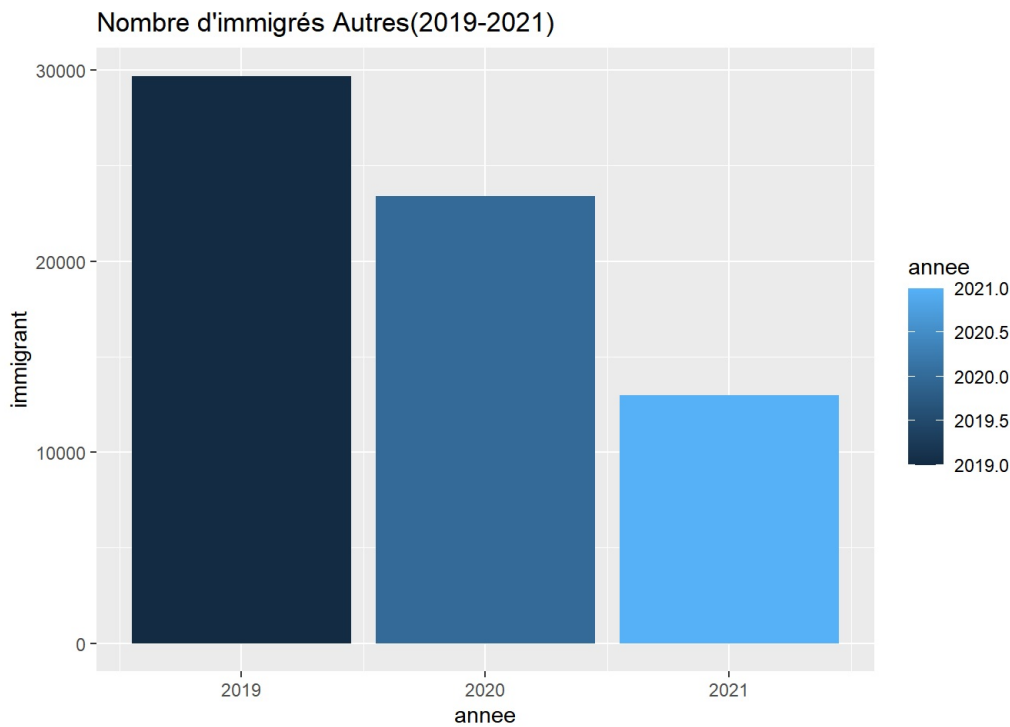
```
ggplot(Asie_immi, aes(x=annee, y=immigrant, fill=annee))+
  geom_col() +
  labs(title="Nombre d'immigrés Asie(2019-2021)")
```

Nombre d'immigrés Asie(2019-2021)



(5) 아메리카, 오세아니아 대륙 연도별 추이

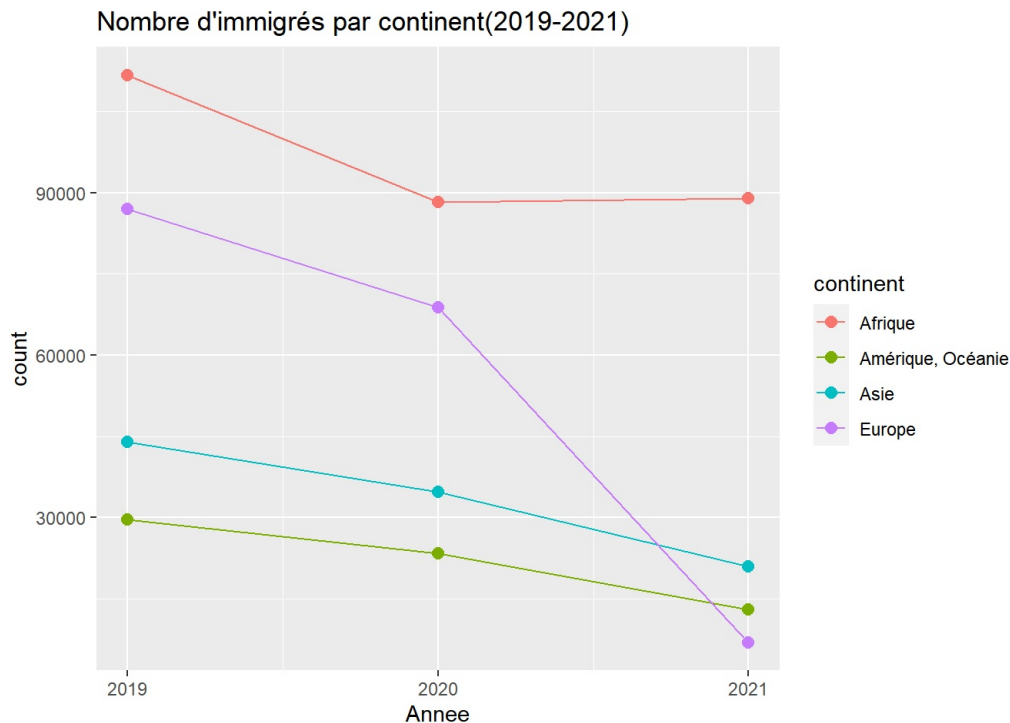
```
ggplot(Amer_immi, aes(x=annee, y=immigrant, fill=annee))+
  geom_col() +
  labs(title="Nombre d'immigrés Autres(2019-2021)") +
  scale_y_continuous(breaks = seq(0,100000,10000))
```

(6) 모든 대륙 비교

모든 대륙의 연도별 추이를 하나의 line chart로 시각해보았다. 그래프를 살펴본 결과 아프리카 대륙 출신을 제외하면 모든 대륙이 하락하는 양상을 띄었고 특히 유럽의 경우 그 하락세가 가파른 것으로 확인되었다.

```
ggplot(immi19_21, aes(x=annee, colour=continent))+
  geom_line(aes(y=immigrant))+
  geom_point(aes(y=immigrant), size=2.5)+
  xlab('Annee')+
  ylab('count')+
  labs(title="Nombre d'immigrés par continent(2019-2021)")+
  scale_x_continuous(breaks = seq(2019,2021,1))
```



4. 프랑스 내 이민자의 거주지역_신지원

4.1. Paris, Marseilles, Lyon 3개 도시 거주 이민자의 출신 대륙

```
df_dep <- read_delim("TCRD_012.txt", delim = "\t",
  escape_double = FALSE, trim_ws = TRUE)
```

```
## New names:
## Rows: 100 Columns: 9
## — Column specification
## _____
## Delimiter: "\t" chr (9): Population immigr?e selon les principaux pays de
## naissance en 2018 ...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## • `` -> `...2`
## • `` -> `...3`
## • `` -> `...4`
## • `` -> `...5`
## • `` -> `...6`
## • `` -> `...7`
## • `` -> `...8`
## • `` -> `...9`
```

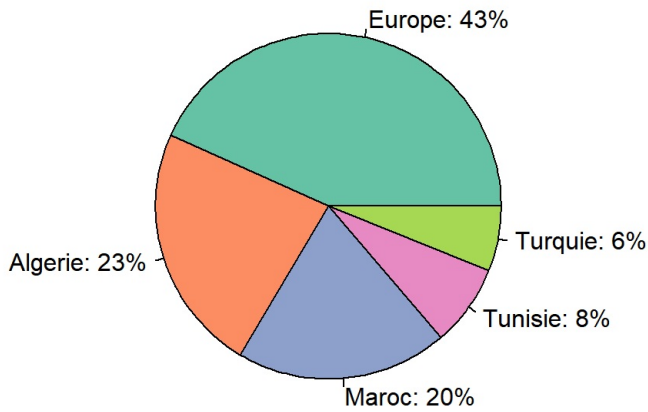
```
df_dep = df_dep[4:100,]
labels = c("Europe", "Algerie", "Maroc", "Tunisie", "Turquie")
```

(1) Paris의 이민자 출신 국가(%)_파이차트

```
Paris = df_dep[76,]
Seine_et_Marne = df_dep[78,]
Yvelines = df_dep[79,]
Paris_deps = df_dep[92:96,]
Paris_region = data.frame(rbind(Paris, Seine_et_Marne, Yvelines, Paris_deps))
Paris_region_Europe_avg = round(mean(x=as.numeric(gsub(",", ".", Paris_region$...5))), na.rm = TRUE),2)
Paris_region_Algerie_avg = round(mean(x=as.numeric(gsub(",", ".", Paris_region$...6))), na.rm = TRUE),2)
Paris_region_Maroc_avg = round(mean(x=as.numeric(gsub(",", ".", Paris_region$...7))), na.rm = TRUE),2)
Paris_region_Tunisie_avg = round(mean(x=as.numeric(gsub(",", ".", Paris_region$...8))), na.rm = TRUE),2)
Paris_region_Turquie_avg = round(mean(x=as.numeric(gsub(",", ".", Paris_region$...9))), na.rm = TRUE),2)
Paris_region_perc = c(Paris_region_Europe_avg, Paris_region_Algerie_avg, Paris_region_Maroc_avg, Paris_region_Tunis
ie_avg, Paris_region_Turquie_avg)
Paris_region_sum = sum(Paris_region_perc)
Paris_region_Europe_avg = round(Paris_region_Europe_avg/Paris_region_sum,2)*100
Paris_region_Algerie_avg = round(Paris_region_Algerie_avg/Paris_region_sum,2)*100
Paris_region_Maroc_avg = round(Paris_region_Maroc_avg/Paris_region_sum,2)*100
Paris_region_Tunisie_avg = round(Paris_region_Tunisie_avg/Paris_region_sum,2)*100
Paris_region_Turquie_avg = round(Paris_region_Turquie_avg/Paris_region_sum,2)*100
Paris_region_perc_2 = c(Paris_region_Europe_avg, Paris_region_Algerie_avg, Paris_region_Maroc_avg, Paris_region_Tun
isie_avg, Paris_region_Turquie_avg)
Paris_region_labels = c(paste(c("Europe: ", Paris_region_Europe_avg, "%"),collapse = ""),
  paste(c("Algerie: ", Paris_region_Algerie_avg, "%"),collapse = ""),
  paste(c("Maroc: ", Paris_region_Maroc_avg, "%"),collapse = ""),
  paste(c("Tunisie: ", Paris_region_Tunisie_avg, "%"),collapse = ""),
  paste(c("Turquie: ", Paris_region_Turquie_avg, "%"),collapse = ""))

pie(Paris_region_perc, Paris_region_labels, main="Paris의 이민자 출신 국가(%)", col=brewer.pal(5,"Set2"))
```

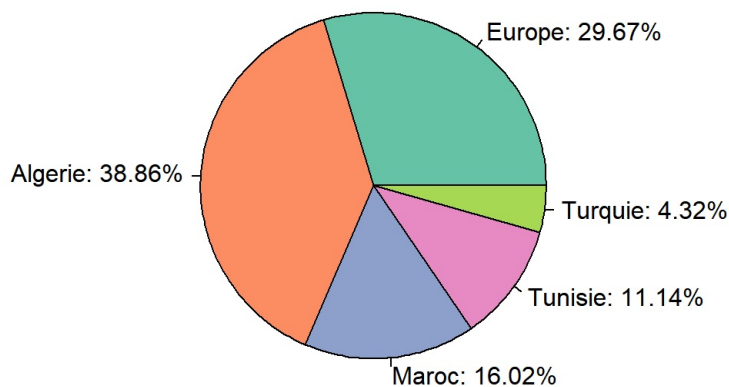
Paris의 이민자 출신 국가(%)



(2) Marseilles의 이민자 출신 국가(%)_파이차트

```
Marseilles = df_dep[11,]
Marseilles_perc = as.numeric(gsub(",", ".", Marseilles[,5:9]))
Marseilles_sum = sum(Marseilles_perc)
Marseilles_Europe_perc = round(Marseilles_perc[1]/Marseilles_sum*100, 2)
Marseilles_Algerie_perc = round(Marseilles_perc[2]/Marseilles_sum*100, 2)
Marseilles_Maroc_perc = round(Marseilles_perc[3]/Marseilles_sum*100, 2)
Marseilles_Tunisie_perc = round(Marseilles_perc[4]/Marseilles_sum*100, 2)
Marseilles_Turquie_perc = round(Marseilles_perc[5]/Marseilles_sum*100, 2)
Marseilles_perc_2 = c(Marseilles_Europe_perc, Marseilles_Algerie_perc, Marseilles_Maroc_perc, Marseilles_Tunisie_perc, Marseilles_Turquie_perc)
Marseilles_labels = c(paste(c("Europe: ", Marseilles_Europe_perc, "%"), collapse = ""),
  paste(c("Algerie: ", Marseilles_Algerie_perc, "%"), collapse = ""),
  paste(c("Maroc: ", Marseilles_Maroc_perc, "%"), collapse = ""),
  paste(c("Tunisie: ", Marseilles_Tunisie_perc, "%"), collapse = ""),
  paste(c("Turquie: ", Marseilles_Turquie_perc, "%"), collapse = ""))
pie(Marseilles_perc_2, Marseilles_labels, main="Marseilles의 이민자 출신 국가(%)", col=brewer.pal(5,"Set2"))
```

Marseilles의 이민자 출신 국가(%)



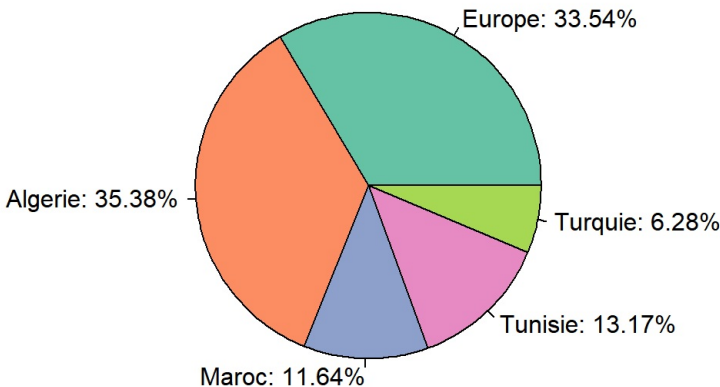
(3) Lyon의 이민자 출신 국가(%)_파이차트

```

Lyon = df_dep[68,]
Lyon_perc = as.numeric(gsub(",", ".", Lyon[,5:9]))
Lyon_sum = sum(Lyon_perc)
Lyon_Europe_perc = round(Lyon_perc[1]/Lyon_sum*100, 2)
Lyon_Algerie_perc = round(Lyon_perc[2]/Lyon_sum*100, 2)
Lyon_Maroc_perc = round(Lyon_perc[3]/Lyon_sum*100, 2)
Lyon_Tunisie_perc = round(Lyon_perc[4]/Lyon_sum*100, 2)
Lyon_Turquie_perc = round(Lyon_perc[5]/Lyon_sum*100, 2)
Lyon_perc_2 = c(Lyon_Europe_perc, Lyon_Algerie_perc, Lyon_Maroc_perc, Lyon_Tunisie_perc, Lyon_Turquie_perc)
Lyon_labels = c(paste(c("Europe: ", Lyon_Europe_perc, "%"),collapse = ""),
                paste(c("Algerie: ", Lyon_Algerie_perc, "%"),collapse = ""),
                paste(c("Maroc: ", Lyon_Maroc_perc, "%"),collapse = ""),
                paste(c("Tunisie: ", Lyon_Tunisie_perc, "%"),collapse = ""),
                paste(c("Turquie: ", Lyon_Turquie_perc, "%"),collapse = ""))
pie(Lyon_perc_2, Lyon_labels, main="Lyon의 이민자 출신 국가(%)", col=brewer.pal(5,"Set2"))

```

Lyon의 이민자 출신 국가(%)



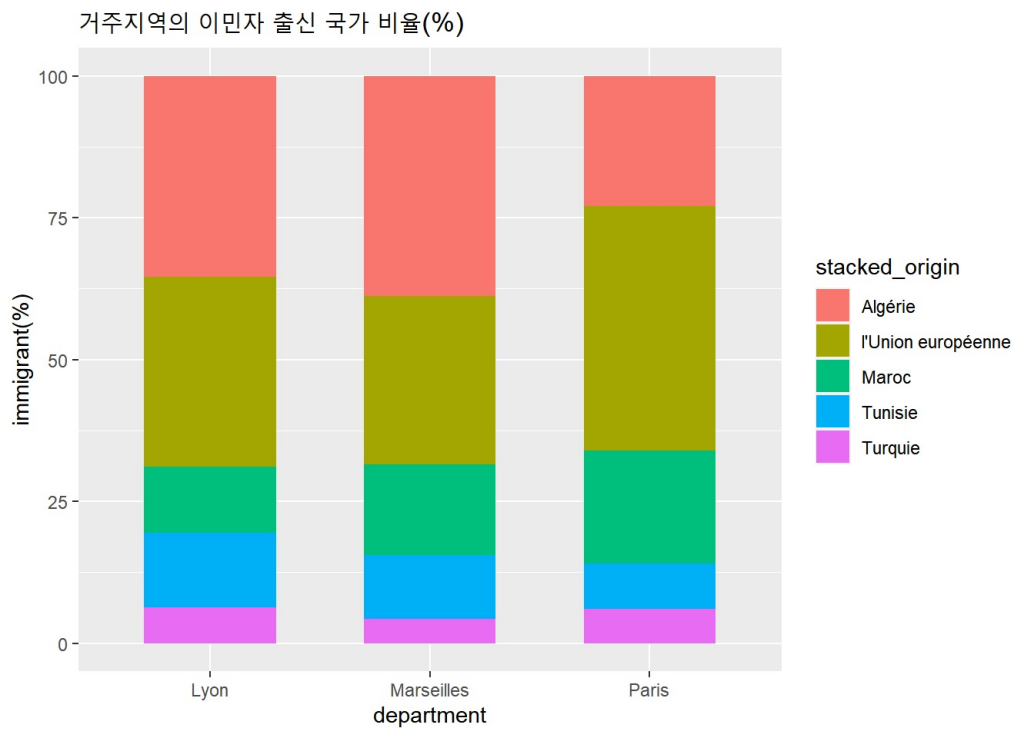
(4) 거주지역의 이민자 출신 국가 비율(%)_누적막대그래프

```

library(ggplot2)
stacked_dep = c(rep(c("Paris", "Marseilles", "Lyon"), each = 5))
stacked_origin = c(rep(c("l'Union européenne", "Algérie", "Maroc", "Tunisie", "Turquie"), times = 3))
stacked_perc = c(43, 23, 20, 8, 6, 29.67, 38.86, 16.02, 11.14, 4.32, 33.54, 35.38, 11.64, 13.17, 6.28)
stacked_data = data.frame(stacked_dep, stacked_origin, stacked_perc)

ggplot(stacked_data, aes(x=stacked_dep, y=stacked_perc, fill=stacked_origin))+
  geom_bar(width =0.6, stat = "identity")+
  labs(title="거주지역의 이민자 출신 국가 비율(%)")+
  labs(x="department",y="immigrant(%)")

```



5. 2019년 프랑스 이민자의 나이 및 학위 -정은아

2018년과 가장 가까운 해인 2019년의 이민자의 전체 성별과 학벌에 관한 데이터를 살펴보도록하겠다.

5.1. 2019년 데이터 불러오기

- 데이터 출처: <https://www.insee.fr/fr/statistiques/3633212#graphique-infographie> (<https://www.insee.fr/fr/statistiques/3633212#graphique-infographie>) 에서 발췌한 csv 파일 내의 Q4 시트를 활용하였다.

```
diplome_age <- read_delim("diplome_age.txt",
                        delim = "\t", escape_double = FALSE,
                        trim_ws = TRUE)
```

```
## Rows: 19 Columns: 6
## — Column specification —————
## Delimiter: "\t"
## chr (1): Caractéristique
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

5.2. 데이터 전처리

```
sex_immi <-diplome_age[1,] #성비 정보만 추출
age_immi <-diplome_age[c(3:7),] #연령 정보만 추출
dip_immi <-diplome_age[c(9:12),] #학위 정보만 추출
```

5.3. 데이터 전처리 및 그래프 생성

(1) 대륙별 이민자의 연령pie chart

```
mycol1 <-c("#FF9E9B", "#FFF56E", "#AAEBAA", "#9BC3FF", "#6E6ED7") # 사용할 색상 저장
opar = par(mfrow=c(2,2), mar=c(2,2,2,2)+0.1)

labels_age <-c("Moins de 18", "18-29", # 나이 범주 입력
              "30-44", "45-64", "65 ou plus")

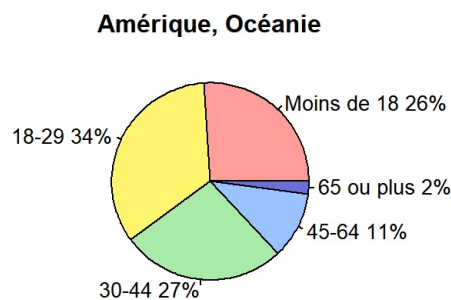
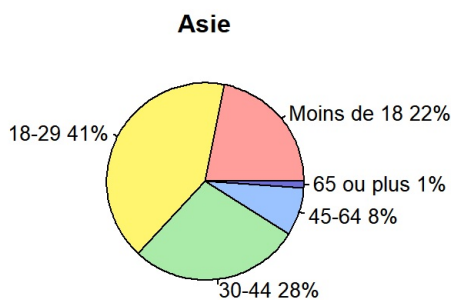
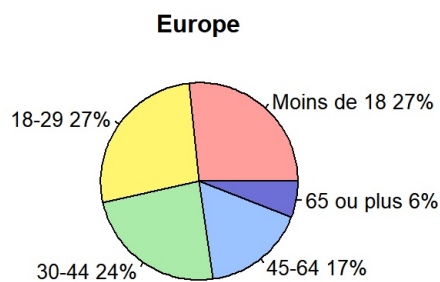
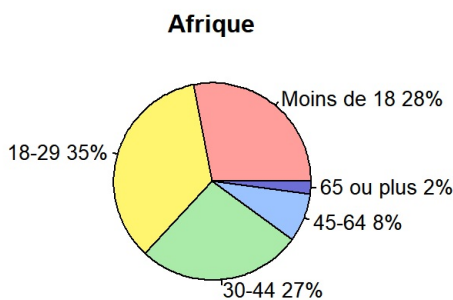
labels_afri <-paste(labels_age, age_immi$Afrique) # 퍼센트 수치 입력
labels_afri <- paste(labels_afri, "%", sep="")

pie(age_immi$Afrique, labels=labels_afri,
    main="Afrique", col =mycol1) # 아프리카 대륙 파이차트 출력

labels_eu <-paste(labels_age, age_immi$Europe)
labels_eu <- paste(labels_eu, "%", sep="")
pie(age_immi$Europe, labels=labels_eu, main="Europe", col =mycol1) #유럽대륙 파이차트

labels_asie <-paste(labels_age, age_immi$Asie)
labels_asie <- paste(labels_asie, "%", sep="")
pie(age_immi$Asie, labels=labels_asie, main="Asie", col =mycol1) #아시아대륙 파이차트

labels_am <-paste(labels_age, age_immi`Amérique, Océanie`)
labels_am <- paste(labels_am, "%", sep="")
pie(age_immi`Amérique, Océanie`, labels=labels_am, main="Amérique, Océanie", col =mycol1) #기타대륙 파이차트
```



시각화 결과 모든 대륙에서 **18세**이서 **29세** 사이의 이민자 비율이 가장 높은 것으로 나타났으며 **65세** 이상의 이민자 수가 가장 낮은 비율을 차지했다.

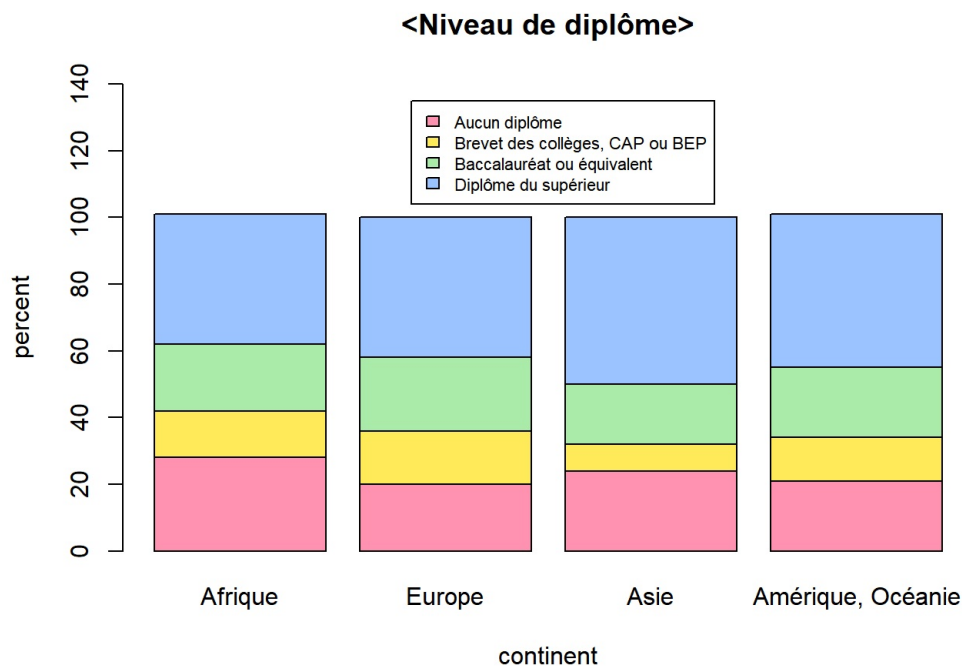
(2) 이민자의 학력수준

전체 이민자의 학력수준을 시각화해보았다. 확인결과 모든 대륙에서 고등교육을 이수한 이민자의 비율이 가장 높은 것으로 나타났다.

```
mycol2 =c("#FF92B1", "#FFEB5A", "#AAEBAA", "#9BC3FF")
dip_count <- as.matrix(dip_immi[,c(2:5)])
opar = par(mfrow=c(1,1))

dip_plot <- barplot(dip_count,
                    main="<Niveau de diplôme>",
                    width=0.5,
                    ylab="percent",
                    xlab="continent",
                    col=mycol2,
                    ylim=c(0,140)) # y축을 140까지 설정한 이유: 범례 추가 공간 확보하기 위해

legend(0.85, 135, dip_immi$Caractéristique,
      fill =mycol2, cex = 0.7)
```



< 3개 도시와 프랑스 전역 데이터 비교 >

데이터 분석에는 마르세유, 리옹, 파리 3개 도시에 대한 데이터를 사용하였다. 분석에 앞서 2018년 3개 도시의 이민자 데이터를 과연 프랑스 전체 이민자의 특성으로 일반화하여도 괜찮은지 알아보기 위해 3개 지역의 이민자의 비율, 성비, 직업분포를 살펴보고 이를 프랑스 전체 지역과 비교했을 때 비슷한 분포를 보이면 3개 도시에 대한 데이터가 프랑스 전역에 관하여 대표성이 있다고 가정하였다.

6. 이민자 비율과 성비 및 직업 범주 비교

6.1. 전체 인구 대비 이민자 비율 비교 -정은아

- 3개 도시의 인구에 대한 데이터를 담은 전처리 파일을 사용하였다.
- <https://www.insee.fr/fr/statistiques/5395871?sommaire=5395920> (<https://www.insee.fr/fr/statistiques/5395871?sommaire=5395920>) 에서 발췌한 데이터를 기반으로 전처리하였다.
- 상세한 전처리 과정은 "데이터전처리3000.csv" 파일을 참고.

```
demo2018 <- read.csv("Demo_count3000.csv", header=T)

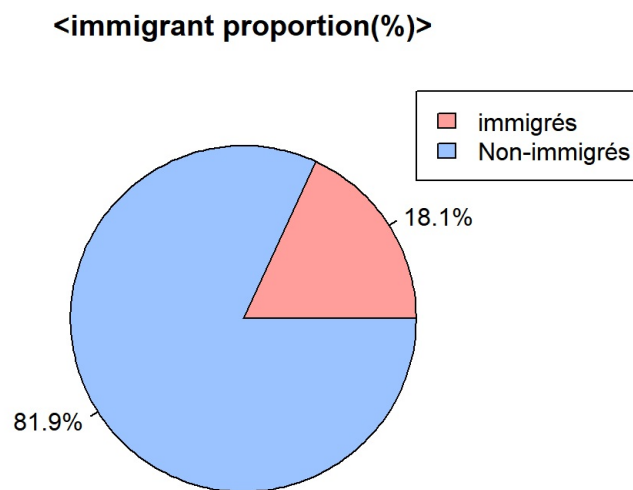
demo2018$immi[demo2018$immi==1] ="immigrés"
demo2018$immi[demo2018$immi==2] ="Non-immigrés"

immi_table <- table(demo2018$immi)
immi_table<-round(prop.table(immi_table)*100,1) # 백분율로 만들어 반올림

immi_labels <- paste(immi_table, "%",sep="")

pie(immi_table,
    labels=immi_labels,
    main="<immigrant proportion(%)>",
    col=c("#FF9E9B", "#9BC3FF"))

immi_labels=c("immigrés", "Non-immigrés")
legend(0.8, 1.05, immi_labels,
    fill = c("#FF9E9B", "#9BC3FF"), cex = 1)
```



* 확인 결과 전체 프랑스 지역의 이민자

의 비율보다 3개 도시에 거주하는 이민자의 비율이 더 높았다. 이같은 결과에 따라 3개 도시의 프랑스 전체 이민자들의 상당수가 거주하고 있는 것으로 판단하여 분석을 프랑스 전체가 아닌 3개도시로 추려서 분석을 진행하게 되었다.

6.2. 이민자의 성비 비교-정은아

(1) 2019년 프랑스 전지역 이민자의 성비


```
# 2019년 프랑스 전체 지역
opar <- par(mfrow=c(1,2))
sex_count <-sex_immi[,c(1,6)]
hommes <- c("Hommes",100-sex_immi[,6]) # 100 - 여성비율 = 남성비율
sex_count[2,]<-hommes # 두번째 열의 남성 데이터 입력

sexe_labels <- paste(sex_count$Caractéristique, sex_count$`Ensemble des immigrés`)
sexe_labels <- paste(sexe_labels, "%",sep="")

pie(sex_count$`Ensemble des immigrés`,
    labels=sexe_labels,
    main="<2019 total region Sexe proportion>",
    col=c("#FF9E9B", "#9BC3FF"),)

# 3개도시 성비 파이차트
demo2018 <- read.csv("Demo_count3000.csv", header=T)
immi_homme <- subset(demo2018, Sexe==1& immi==1) # 남성 이민자
immi_femme <- subset(demo2018, Sexe==2& immi==1) # 여성 이민자
total_sexe <- rbind(immi_homme,immi_femme)

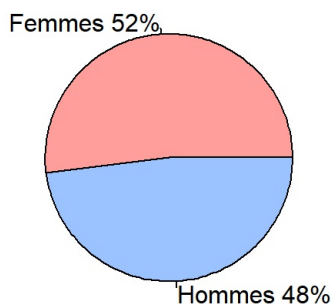
# * 성비를 나타내는 파이차트 생성
total_sexe$Sexe[total_sexe$Sexe==1] ="Homme"
total_sexe$Sexe[total_sexe$Sexe==2] ="Femme"

sexe_table<-table(total_sexe$Sexe)
sexe_table<-round(prop.table(sexe_table)*100,1) # 백분율로 만들어 반올림
sexe_table <-as.matrix(sexe_table)

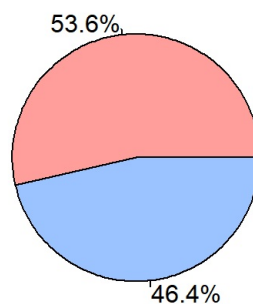
sexe_labels <- paste(sexe_table, "%",sep="")

pie(sexe_table,
    labels=sexe_labels,
    main="<2018 3 region Sexe proportion>",
    col=c("#FF9E9B", "#9BC3FF"))
```

<2019 total region Sexe proportion



<2018 3 region Sexe proportion>



확인 결과 전체 이민자에 비해 3개도시의 성비에서 여성의 비율이 조금 더 높게 나타났으나 여성 53.6% 남성 약 46.4% 비율로 비슷한 수준인 것으로 확인되었다. 따라서 해당 데이터가 대표성에 문제가 없는 것으로 가정하였다.

6.3. 직업 분포 비교

(1) 2018년 프랑스 전체지역 인구의 직업 분포 - 지명은

이민자의 노동시장 현황에 관련한 통계분석에 사용할 데이터는 2018년도의 데이터이다. 때문에 2018년 프랑스의 이민자와 비이민자가 주로 종사하였던 직종이 어떤 것인지 현황을 파악해보고 시각화해보고자한다.

• 2018년도 데이터 불러오기

```
domaines2018 <- read.csv("immigrés_secteurs_metiers_2018.csv", encoding = "UTF-8", header = T)
```

- 데이터 정제하기 먼저 비이민자가 주로 종사하는 직업 상위 5부문을 추출하여 데이터를 정제하였다.

```
# 비이민자의 상위 5개 직업 추출 및 정제
idp <- domaines2018[, c(2:4)]
idp <- domaines2018[c(4, 9, 18, 22, 29, 33, 40, 44, 46, 53, 55, 63, 67, 69, 75, 79, 85, 90, 98, 101, 108, 111), c(2:4)]

names(idp) <- c("domaine", "nonimmigrés", "immigrés")

# 상위 5개 직업 분야 추출
NI2018 <- idp[,c(1,2)]
NI2018 <- idp[c(11,14,16,18,20), c(1,2)]

# 자료 타입 변환
NI2018$nonimmigrés <- gsub("","",NI2018$nonimmigrés) # 숫자 쉼표 제거
NI2018$nonimmigrés <- as.numeric(NI2018$nonimmigrés) # 숫자형 자료로 변환

# 이름 간소화
domaine1 <- c("Gestion", "Administration", "Commerce", "Services", "Santé")
NI2018$domaine <- domaine1
```

- 비이민자의 상위 5개 직업 추출 및 정제 다음으로는 이민자가 주로 종사하는 직종 상위 5개 부문을 추출하여 데이터를 정제하였다.

```
# 상위 5개 직업 분야 추출
I2018 <- idp[,c(1,3)]
I2018 <- idp[c(2,9,11,16,18),c(1,3)]

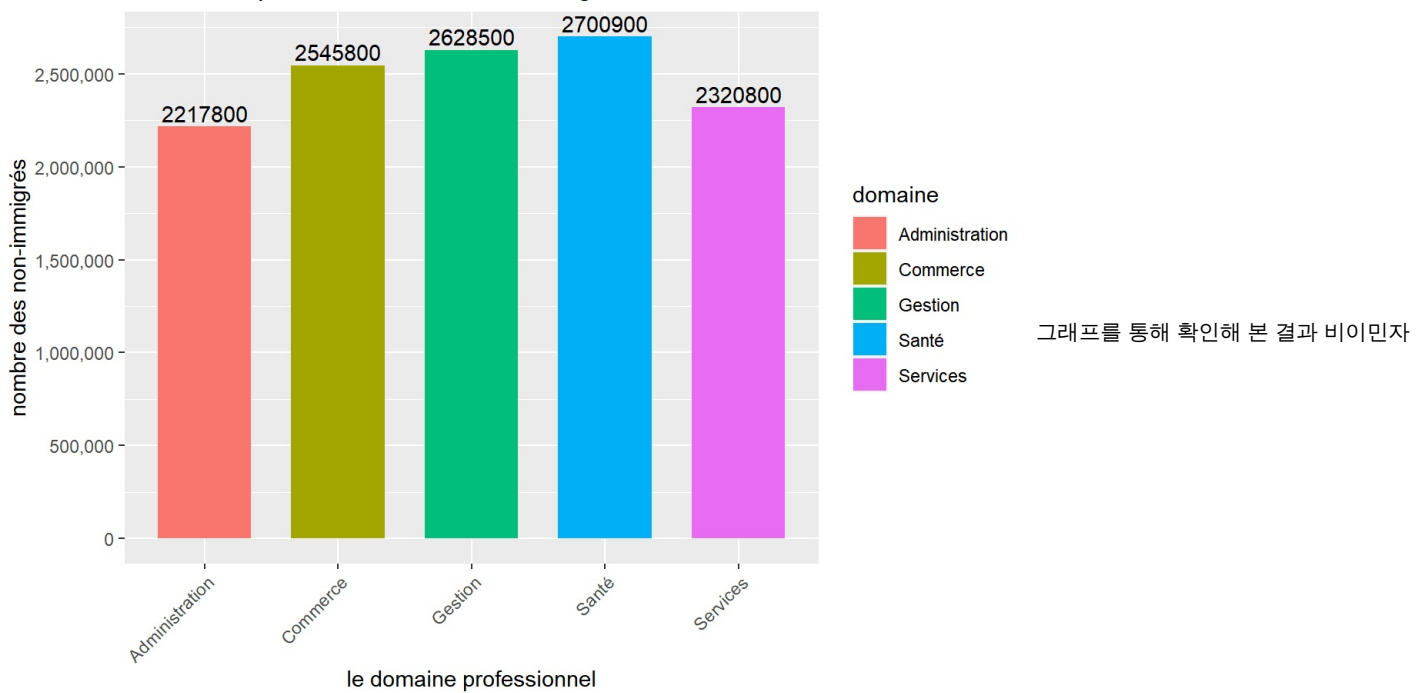
I2018$immigrés <- gsub("","",I2018$immigrés) # 숫자 쉼표 제거
I2018$immigrés <- as.numeric(I2018$immigrés) # 숫자형 자료로 변환

# 이름 간소화
domaine2 <- c("Bâtiment", "Transports", "Gestion", "Commerce", "Services")
I2018$domaine <- domaine2
```

- 시각화 하기

```
# 비이민자 상위 5개 직업 분야
ggplot(NI2018, aes(x=domaine, y=nonimmigrés, fill=domaine))+
  geom_bar(width = 0.7, stat="identity" )+
  labs (x="le domaine professionnel",
        y="nombre des non-immigrés",
        title="le domaine professionnel des non-immigrés en 2018")+
  theme(axis.text.x = element_text(angle=45, hjust=1, size=8.5))+
  geom_text(aes(label=nonimmigrés), vjust=-0.3) +
  scale_y_continuous(labels = scales::comma, breaks = seq(0,3000000,500000))
```

le domaine professionnel des non-immigrés en 2018

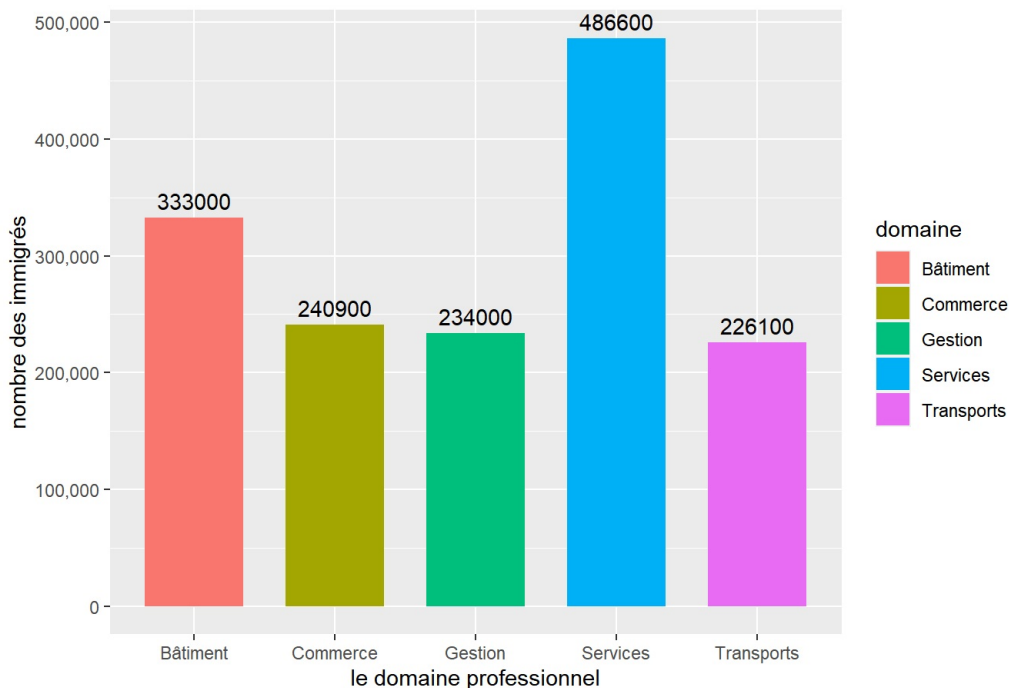


의 경우 **Santé, action sociale, culturelle et sportive(의료, 문화 및 스포츠)** 직종에 종사하는 사람이 가장 많았다. 여기에는 보건위생분야에 종사하는 의사와 간병인, 그리고 스포츠 관련 종사자를 포함한다. 다음으로는 **Gestion, administration des entreprises(경영)** 종사하는 사람이 많았다. 여기에는 비서나, 회계업무를 맡는 직원, 비즈니스 리더 등 경영분야에 종사하는 사람들이 포함되었다. 세번째로 가장 많은 부문을 차지한 것은 **Commerce(상업)**에 종사하는 사람이었다. 여기에는 판매원, 영업담당자 등이 이에 속한다.

- 이민자의 상위 5개 직업 추출 및 정제 다음으로는 이민자가 종사하고 있는 직업 상위 5개 분야를 시각화하였다.

```
# 이민자 상위 5개 직업 분야
ggplot(I2018, aes(x=domaine, y=immigrés, fill=domaine))+
  geom_bar(width = 0.7, stat="identity")+
  labs (x="le domaine professionnel",
        y="nombre des immigrés",
        title="le domaine professionnel des immigrés en 2018")+
  geom_text(aes(label=immigrés), vjust=-0.5) +
  scale_y_continuous(labels = scales::comma, breaks = seq(0,700000,100000))
```

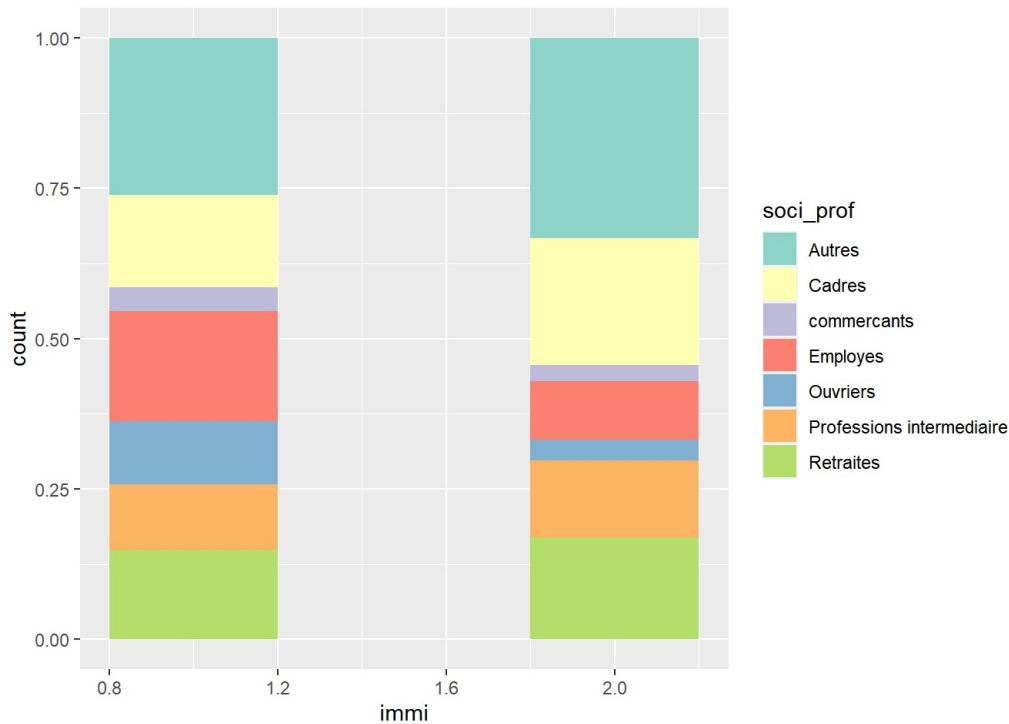
le domaine professionnel des immigrés en 2018



눈에 띄는 것은 비이민자의 직업 분포에서 4위를 차지하였던 **Services aux particuliers et aux collectivités (서비스업 종사자)**가 가장 많은 비율을 차지하고 있다는 점이다. 그 다음순위를 차지하는 **Bâtiment, travaux publics(건축업)** 또한 비이민자의 직업부문 5순위안에 들지 않았던 분야이다. 마지막으로 주목할 만한 점은 비이민자 직업 상위 5위에 없었던 **Transports, logistique et tourisme (교통 물류업 및 관광업)** 분야가 이민자 집단에서는 5위로 올라와있다는 점이다.

(2) 3개 도시 인구의 직업 분포-정은아

```
ggplot(demo2018, aes(x=immi, fill=soci_prof)) +
  geom_bar(width=0.4, position='fill') +
  scale_fill_brewer(palette="Set3")
```



3개도시 거주 인구의 직업 분포를 이민여부에 따라 확인해본 결과 2018년 프랑스 전체에 직업 분포와 3개도시의 직업분포가 다른 범주를 사용하고 있지만 이민자 중 가장 많은 사람이 종사하던 서비스직종과 건축업이 각각 employés(직원)과 ouvrier(단순노동직)에 포함되어있는 직종이었다. 확인해보면 **employés**의 경우 **autres**(비경제활동)과 **retraites**(은퇴)를 제외하고 직업 중 1순위를 차지하여 프랑스 전체 직업 분포와 3개도시의 분포가 일치하는 것을 확인할 수 있었다.

또한 이민자가 비이민자보다 상대적으로 많이 분포하고 있는 직업군은 **ouvrier**(단순노동직)으로 해당 범주에는 건축업 현장 노동자, 운송업자, 자동차 정비사 등의 직업군이 분포한다. 또 이민자에 비해 비이민자의 비율이 상대적으로 높았던 직업은 **Cadres**(전문직)으로 해당 직종에는 변호사, 의사, 각 종기관의 임원 등의 직업군이 분포하였다.

7. 교차분석

7.1. 직업범주와 이민여부의 연관성-지명은

(1) 데이터 정제

```
#### (1) 패키지 설치 및 로딩
#install.packages(("gmodels"))
library(gmodels)

#### (2) 이민자 직업 데이터 읽기
immiprof <- read.csv("Demo_count3000.csv", stringsAsFactors = T)

#### (3) 요약 통계 확인
summary(immiprof)
```

```
##           X           ID           Sexe           immi
## Min.      : 1.0      Min.      : 1.0      Min.      :1.00      Min.      :1.000
## 1st Qu.: 891.5      1st Qu.: 891.5      1st Qu.:1.00      1st Qu.:2.000
## Median :1782.0      Median :1782.0      Median :2.00      Median :2.000
## Mean     :1782.0      Mean     :1782.0      Mean     :1.53      Mean     :1.819
## 3rd Qu.:2672.5      3rd Qu.:2672.5      3rd Qu.:2.00      3rd Qu.:2.000
## Max.     :3563.0      Max.     :3563.0      Max.     :2.00      Max.     :2.000
##
##                soci_prof      region
## Autres           :1140      Lyon : 519
## Cadres           : 715      Mar  : 868
## commercants      : 104      Paris:2176
## Employes         : 403
## Ouvriers         : 167
## Professions intermediaire: 442
## Retraites        : 592
```

```
#### (4) 범주형 데이터 정제
immiprof$immi <- factor(immiprof$immi, levels = c(1:2),
                        labels=c('immigrant', 'non_immigrant'))
```

```
#### (5) 정제 결과를 요약 통계로 확인
summary(immiprof)
```

```
##           X           ID           Sexe           immi
## Min.      : 1.0    Min.      : 1.0    Min.      :1.00    immigrant      : 645
## 1st Qu.: 891.5    1st Qu.: 891.5    1st Qu.:1.00    non_immigrant:2918
## Median :1782.0    Median :1782.0    Median :2.00
## Mean      :1782.0    Mean      :1782.0    Mean      :1.53
## 3rd Qu.:2672.5    3rd Qu.:2672.5    3rd Qu.:2.00
## Max.      :3563.0    Max.      :3563.0    Max.      :2.00
##
##           soci_prof           region
## Autres              :1140    Lyon : 519
## Cadres              : 715    Mar  : 868
## commercants        : 104    Paris:2176
## Employes           : 403
## Ouvriers           : 167
## Professions intermediaire: 442
## Retraites          : 592
```

(2) 교차분석 수행

'이민여부'에 따라 '직업'는 유의미한 관계를 갖는가

```
CrossTable(immiprof$soci_prof, immiprof$immi, chisq = T, expected = T)
```

```
##
##
## Cell Contents
## |-----|
## |              N |
## |      Expected N |
## | Chi-square contribution |
## |      N / Row Total |
## |      N / Col Total |
## |      N / Table Total |
## |-----|
##
##
## Total Observations in Table:  3563
##
##
##           immiprof$soci_prof | immiprof$immi
##           immigrant | non_immigrant | Row Total |
## -----|-----|-----|-----|
##           Autres |          168 |          972 |          1140 |
##           |          206.371 |          933.629 |          |
##           |          7.134 |          1.577 |          |
##           |          0.147 |          0.853 |          0.320 |
##           |          0.260 |          0.333 |          |
##           |          0.047 |          0.273 |          |
## -----|-----|-----|-----|
##           Cadres |          99 |          616 |          715 |
##           |         129.434 |          585.566 |          |
##           |          7.156 |          1.582 |          |
##           |          0.138 |          0.862 |          0.201 |
##           |          0.153 |          0.211 |          |
##           |          0.028 |          0.173 |          |
## -----|-----|-----|-----|
##           commercants |          26 |          78 |          104 |
##           |         18.827 |          85.173 |          |
##           |          2.733 |          0.604 |          |
##           |          0.250 |          0.750 |          0.029 |
##           |          0.040 |          0.027 |          |
##           |          0.007 |          0.022 |          |
## -----|-----|-----|-----|
##           Employes |         118 |          285 |          403 |
##           |         72.954 |          330.046 |          |
##           |         27.814 |          6.148 |          |
##           |          0.293 |          0.707 |          0.113 |
##           |          0.183 |          0.098 |          |
##           |          0.033 |          0.080 |          |
## -----|-----|-----|-----|
```

```
##          Ouvriers |          68 |          99 |          167 |
##          30.232 |        136.768 |
##          47.184 |         10.430 |
##          0.407 |         0.593 |         0.047 |
##          0.105 |         0.034 |
##          0.019 |         0.028 |
## -----|-----|-----|-----|
## Professions intermediaire |          70 |          372 |          442 |
##          80.014 |        361.986 |
##          1.253 |         0.277 |
##          0.158 |         0.842 |         0.124 |
##          0.109 |         0.127 |
##          0.020 |         0.104 |
## -----|-----|-----|-----|
##          Retraites |          96 |          496 |          592 |
##          107.168 |        484.832 |
##          1.164 |         0.257 |
##          0.162 |         0.838 |         0.166 |
##          0.149 |         0.170 |
##          0.027 |         0.139 |
## -----|-----|-----|-----|
##          Column Total |          645 |          2918 |          3563 |
##          0.181 |         0.819 |
## -----|-----|-----|-----|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## -----
## Chi^2 = 115.3142      d.f. = 6      p = 1.568862e-22
##
##
##
```

#####1단계: 변수 간의 연관성이 통계적으로 유의한지 여부 확인

p-value (1.568862e-22)는 유의수준 0.05보다 작은 값이므로 귀무가설은 귀각, 대립가설은 채택 0.05의 유의 수준은 변수 간에 실제로 연관성이 없는데 연관성이 존재한다는 결론을 내릴 위험이 5%라는 것을 나타낸다. (*귀무가설: 이민여부에 따른 직업은 관계가 없다.* 대립가설: 이민여부에 따른 직업은 유의미한 관계를 갖는다.) (p-값 ≤ 0.05: 변수 간에 통계적으로 유의한 연관성이 있음) p-값이 유의 수준보다 작거나 같으면 귀무 가설을 기각하고 변수 간에 통계적으로 유의한 연관성이 있다는 결론을 내림 따라서 이민여부에 따른 직업은 유의미한 관계를 갖는다(영향을 미친다/차이를 갖는다). -> 이민 여부에 따라 직업의 차이가 존재

#####2단계: 기대 카운트와 관측 카운트 간의 차이를 조사하여 연관성에 가장 크게 영향을 미치는 변수 수준 확인 Chi-square Contribution값이 가장 큰 경우는 soci_prof = Ouvriers(단순 노동직) & immi = immigrant 이 셀의 관측빈도는 68 이며, 기대빈도는 30.232 이것을 정리해보면, 두 변수가 독립임을 가정했을 때 빈도가 30명 정도로 기대된 것에 반해 실제 조사된 관측빈도는 훨씬 많은 68명이 나온 것이다. (기대빈도: 변수가 서로 독립적인 경우 하나의 셀에서 평균적으로 기대되는 빈도, 관측빈도: 표본에서 한 범주에 속하는 관측치의 실제 수) (이 결과에서 관측된 셀 카운트는 각 셀의 첫 번째 숫자이고, 기대 카운트는 각 셀의 두 번째 숫자임) 결과를 보면, 단순 노동직에 종사하는 그룹에 이민자의 실제 수가 기대값보다 많음을 알 수 있다. 반대로 soci_prof = Ouvriers(단순 노동직) & immi = non_immigrant셀을 보면 관측빈도는 99, 기대빈도는 136.768 이는 단순 노동직에서 종사하는 비이민자의 수는 기대값보다 훨씬 더 적음을 뜻한다. 따라서 단순 노동직 부분에서 비이민자의 수는 교차분석에서 기대되는 값보다 적었고, 이민자의 수는 교차분석에서 기대되는 값보다 많았다. => '이민여부'에 따라 '직업'은 관계가 있다.

(3) 이민자 수와 비이민자 수의 직업별 분포도 시각화하기

‘이민경험’ 여부 있음에 대한 분석

```
library(ggplot2)
#### (7)-1. 벡터 설정하기

i <- c("기대값", "관측값")
soci_prof <- c("Autres", "Cadres", "commerçants", "Employes", "Ouvriers", "professions intermediaire", "Retraites")

#### (7)-2. 데이터 프레임 만들기

df <- expand.grid(soci_prof, i)
df$value <- c(206.371, 129.434, 18.827, 72.954, 30.232, 80.014, 107.168, 168, 99, 26, 118, 68, 70, 96)
g <- ggplot(df, aes(Var2, Var1)) + geom_point(aes(size = value), colour = "Orange") + theme_bw() + xlab("") + ylab("")
g + scale_size_continuous(range=c(8,24)) + geom_text(aes(label = value))
```



‘이민경험’ 여부 없음에 대한 분석

(7)-1. 벡터 설정하기

```
ni <- c("기대값", "관측값")
soci_prof <- c("Autres", "Cadres", "commercants", "Employes", "Ouvriers", "professions intermediaire", "Retraites")
```

(7)-2. 데이터 프레임 만들기

```
df <- expand.grid(soci_prof, ni)
df$value <- c(933.629, 585.566, 85.173, 330.046, 136.768, 361.986, 484.832, 972, 616, 78, 285, 99, 372, 496)

g <- ggplot(df, aes(Var2, Var1)) + geom_point(aes(size = value), colour = "skyblue") + theme_bw() + xlab("") + ylab("")
g + scale_size_continuous(range=c(8,24)) + geom_text(aes(label = value))
```



이민여부에 따라 직업은 관계가 있는 지에 대한 교차분석을 토대로 시각화해보았다. 위는 이민자에 대한 직업분포와 비이민자에 대한 직업분포의 기대값과 관측값을 각각 나타낸 것이다. 두 시각화 그래프를 비교해 보았을 때, 가장 눈에 띄는 차이점은 Ouvriers(단순 노동직)와 commercants(자영업) 분야이다. 이 직업 분야에 대해 'immigrant'의 경우 기대값보다 관측값이 더 크게 나왔지만, 'non_immigrant'의 경우 더 작게 나온 것으로 확인된다. 단순 노동직 분야에서는 기대보다 더 적은 비이민자가 종사하며, 기대보다 더 많은 이민자가 종사했다. -> '이민여부'에 따라 '직업' 분야는 관련이 있다.

7.2. 이민여부와 지역의 관계_교차분석_신지원

(1) 데이터 전처리

```
#### (1) 패키지 설치 및 로딩
#install.packages(("gmodels"))
library(gmodels)

#### (2) 2018 전처리 데이터 읽기
immiregion <- read.csv("Demo_count3000.csv", stringsAsFactors = T)

immiregion$immi <- factor(immiregion$immi, levels=c(1:2),
                        labels=c('immigrant', 'non_immi'))

#### (3) 요약통계 확인
summary(immiregion)
```

##	X	ID	Sexe	immi
##	Min. : 1.0	Min. : 1.0	Min. :1.00	immigrant: 645
##	1st Qu.: 891.5	1st Qu.: 891.5	1st Qu.:1.00	non_immi :2918
##	Median :1782.0	Median :1782.0	Median :2.00	
##	Mean :1782.0	Mean :1782.0	Mean :1.53	
##	3rd Qu.:2672.5	3rd Qu.:2672.5	3rd Qu.:2.00	
##	Max. :3563.0	Max. :3563.0	Max. :2.00	
##				
##		soci_prof	region	
##	Autres	:1140	Lyon : 519	
##	Cadres	: 715	Mar : 868	
##	commerçants	: 104	Paris:2176	
##	Employes	: 403		
##	Ouvriers	: 167		
##	Professions intermédiaire	: 442		
##	Retraites	: 592		

(2) 교차분석 수행

```
CrossTable(immiregion$region, immiregion$immi, chisq=T, expected=T)
```



```
##
##
## Cell Contents
## |-----|
## |                N |
## | Expected N |
## | Chi-square contribution |
## | N / Row Total |
## | N / Col Total |
## | N / Table Total |
## |-----|
##
##
## Total Observations in Table: 3563
##
##
##
## | immiregion$immi
## |
## | immigrant | non_immi | Row Total |
## |-----|-----|-----|
## | Lyon      | 69       | 450      | 519      |
## |           | 93.953   | 425.047  |           |
## |           | 6.627    | 1.465    |           |
## |           | 0.133    | 0.867    | 0.146    |
## |           | 0.107    | 0.154    |           |
## |           | 0.019    | 0.126    |           |
## |-----|-----|-----|
## | Mar       | 134      | 734      | 868      |
## |           | 157.132  | 710.868  |           |
## |           | 3.405    | 0.753    |           |
## |           | 0.154    | 0.846    | 0.244    |
## |           | 0.208    | 0.252    |           |
## |           | 0.038    | 0.206    |           |
## |-----|-----|-----|
## | Paris     | 442      | 1734     | 2176     |
## |           | 393.915  | 1782.085 |           |
## |           | 5.870    | 1.297    |           |
## |           | 0.203    | 0.797    | 0.611    |
## |           | 0.685    | 0.594    |           |
## |           | 0.124    | 0.487    |           |
## |-----|-----|-----|
## | Column Total | 645     | 2918     | 3563     |
## |           | 0.181    | 0.819    |           |
## |-----|-----|-----|
##
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## -----
## Chi^2 = 19.41729    d.f. = 2    p = 6.075602e-05
##
##
##
```

1단계: 변수 간의 연관성이 통계적으로 유의한지 여부 확인

- 귀무가설: 이민여부에 따른 지역은 관계가 없다.
- 대립가설: 이민여부에 따른 지역은 유의미한 관계가 있다. $\#(p\text{-값} \leq 0.05: \text{변수 간에 통계적으로 유의한 연관성이 있음})$ $\#(p\text{-value } (6.075602e-05) \text{ 는 유의수준 } 0.05 \text{보다 작은 값이므로 귀무가설은 귀각, 대립가설은 채택})$ (0.05의 유의 수준은 변수 간에 실제로 연관성이 없는데 연관성이 존재한다는 결론을 내릴 위험이 5%라는 것을 나타낸다.)
- 결과: 따라서 이민여부에 따른 지역은 유의미한 관계를 갖는다(영향을 미친다/차이를 갖는다). -> 이민여부에 따라 지역의 차이가 존재

2단계: 기대 카운트와 관측 카운트 간의 차이를 조사하여 연관성에 가장 크게 영향을 미치는 변수 수준 확인

$\#(\text{기대빈도: 변수가 서로 독립적인 경우 하나의 셀에서 평균적으로 기대되는 빈도, 관측빈도: 표본에서 한 범주에 속하는 관측치의 실제 수})$ $\#(\text{첫번째 셀: 관측값 / 두번째 셀: 기대값})$

Chi-square Contribution값이 가장 큰 경우는 `immiregion\region = immiregion\immigrant = Lyon` 이 셀의 관측빈도는 69이며, 기대빈도는 93.953 정리해보면, 빈도가 94명 정도로 기대된 것에 반해 실제 조사된 관측빈도는 69명이 나온 것이다. Lyon과 Marseilles는 기대된 것에 비해 적은 이민자가 유입됐으나, Paris는 기대된 것보다 더 많은 이민자가 유입됨을 알 수 있다.

반대로 `immiregion\region = Lyon & immiregion\region = non_immigrant` 셀을 보면 관측빈도는 450, 기대빈도는 425.047 이는 Lyon에서 기대된 것보다 비이민자가 더 많다는 것을 뜻한다. Lyon과 Marseilles에 기대된 것에 비해 많은 비이민자가 거주했으나, Paris는 기대된 것보다 더 적은 비이민자가 거주했음을 알 수 있다. Paris는 기대보다 더 많은 이민자가 유입됐으며, 더 적은 비이민자가 거주했다. 반면 Lyon과 Marseilles에서는 기대보다 더 적은 이민자가 유입됐으며, 더 많은 비이민자가 거주했다.=> '이민여부'에 따라 '지역'은 관계가 있다.

(3) 분석 결과 시각화

```
library(ggplot2)
```

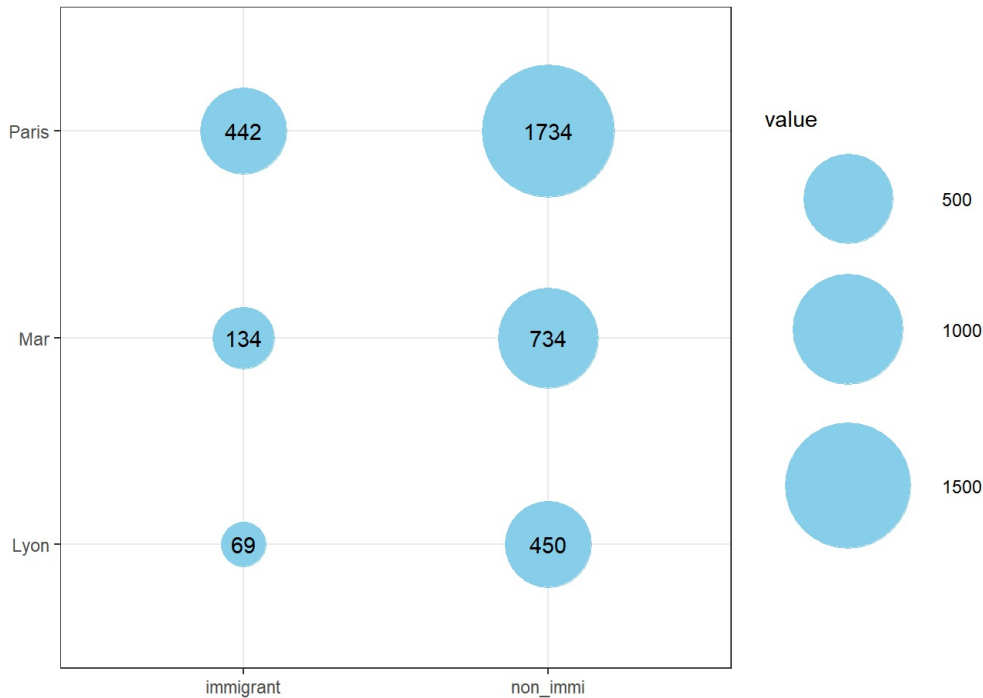
```
#### (5)-1. 벡터 설정하기
```

```
immi <- c("immigrant","non_immi")  
region <- c("Lyon", "Mar", "Paris")
```

```
#### (5)-2. 데이터 프레임 만들기
```

```
df <- expand.grid(region, immi)  
df$value <- c(69, 134, 442, 450, 734, 1734)
```

```
g <- ggplot(df, aes(Var2, Var1)) + geom_point(aes(size = value), colour = "Skyblue") + theme_bw() + xlab("") + ylab("")  
g + scale_size_continuous(range=c(10,30)) + geom_text(aes(label = value))
```



Paris의 경우, 이민자 대비 비이민자 비율이 3.92:1로, 대략 비이민자 4명 당 이민자 1명이 있다고 해석할 수 있다.

Marseilles의 경우, 이민자 대비 비이민자 비율은 5.48:1로, 대략 비이민자 5.5명 당 이민자 1명이 있다고 해석할 수 있다.

Lyon의 경우, 이민자 대비 비이민자 비율은 6.52:1로, 대략 비이민자 6.5명 당 이민자 1명이 있다고 해석할 수 있다.

따라서 Paris, Marseilles, Lyon 순으로 이민자와 비이민자의 비율이 높게 나타난 것으로 보아 이민 여부와 지역의 관계가 유의미하다는 것을 알 수 있다.

8. 로지스틱 회귀분석 - 정은아

8.0. 모델 데이터 로딩

모델의 크기가 무겁기때문에 미리 모델을 저장해놓은 rds 파일을 로딩해주었다.

```
# 통계 분석 모델 로딩
```

```
model1_2 <- readRDS("./model1_2.rds")  
model2_2 <- readRDS("./model2_2.rds")  
model3_2 <- readRDS("./model3_2.rds")  
model4_2 <- readRDS("./model4_2.rds")  
model5_2 <- readRDS("./model5_2.rds")  
model6_2 <- readRDS("./model6_2.rds")  
model7_2 <- readRDS("./model7_2.rds")
```

8.1. 데이터 불러오기

- 데이터 설명: 먼저 마르세유, 파리, 리옹에 거주하는 356만명에 대한 인구 데이터를 1000대 1로 축소 샘플링한 전처리 데이터를 불러온다. 해당 데이터는 총 3563개의 데이터로 이루어져 있다.
- 출처: 해당 데이터는 "<https://www.insee.fr/fr/statistiques/5395871?sommaire=5395920>"에서 (<https://www.insee.fr/fr/statistiques/5395871?sommaire=5395920%22>에서) 발췌한"MG3A – Population par sexe, situation quant à l'immigration et catégorie socioprofessionnelle" 자료를 활용하였다.

```
immi2018<- read.csv("Demo_count3000.csv", encoding="UTF-8")  
immi2018 <- immi2018[, -1]
```

8.2. 분석 전 데이터 정제

로지스틱 회귀분석에서는 데이터의 형태가 범주형이면서 0과 1로 이루어진 이진수로 구성되어있어야 한다. 따라서 범주형 데이터를 0과 1로 재구성해 주었다.

(1) 숫자형으로 변환

```
immi2018$Sexe <- as.numeric(immi2018$Sexe)
immi2018$immi <- as.numeric(immi2018$immi)
```

(2) 이민과 성별 데이터 이진형태로 변환

```
immi2018$Sexe <- ifelse(immi2018$Sexe==1,"Homme","Femme")
immi2018$immi <- ifelse(immi2018$immi==1,"immigr  ","Non-immigr  ")
```

(3) 각 직업 여부를 나타낸 이진수로 변환

```
immi2018$prof_autres<-ifelse(immi2018$soci_prof== "Autres",1,0) #비경제활동 여부
immi2018$prof_commer<-ifelse(immi2018$soci_prof== "commercants",1,0) # 상업 종사자 여부
immi2018$prof_Cadres<-ifelse(immi2018$soci_prof== "Cadres",1,0) #전문직 종사 여부
immi2018$prof_Profess<-ifelse(immi2018$soci_prof== "Professions intermediaire",1,0) # 중간직 종사 여부
immi2018$prof_Employes<-ifelse(immi2018$soci_prof== "Employes",1,0) # 직원 및 서비스업 종사 여부
immi2018$prof_Ouvriers<-ifelse(immi2018$soci_prof== "Ouvriers",1,0) # 단순노동직 종사 여부
immi2018$prof_Retra<-ifelse(immi2018$soci_prof== "Retraites",1,0) # 은퇴 여부
```

(4) 지역 기준 설정

```
immi2018$region = factor(immi2018$region , levels = c("Paris","Mar","Lyon"))
# 파리 지방을 가장 첫번째 레벨로 지정하여 파리 변수 기준으로 모델이 구축되도록 함
```

8.3. 로지스틱 회귀분석

- 정의: 로지스틱 회귀분석이란 독립변수가 종속변수에 영향을 미치는지 아닌지를 파악할 수 있는 통계 분석 기법 이다. 독립변수와 종속변수의 정보가 포함된 데이터를 모델에 학습시켰을때 특정 독립변수를 가지고 있을 때 종속변수가 어떻게 변하는지 예측해준다.
- 활용:로지스틱 회귀분석은 고객의 정보를 기반으로한 보험사기 여부 예측, 고객의 구매 가능성 예측, 질병의 양성/음성 여부 예측하는 데에 사용 된다.
- 분석 내용: 로지스틱 회귀분석 모델에 입력되는 종속변수는 범주형이면서 이진수 형태로 이루어져있어야한다. 우리는 해당 분석을 통해 프랑스 인구의 성별, 거주지, 이민여부를 기반으로 어떠한 인구 정보를 가지고 있을때 어떤 직종에 종사할 확률이 높아지는지 예측해볼 수 있다. 해당 분석을 통해 이민여부가 특정 직업 종사 여부에 어떤 방향으로 작용하고 있는지 파악할 수 있다.
- (참고) 분석에 사용된 직업 분류는 총 8개 범주였다. 하지만 농업에 종사하는 인구가 3개도시에 1000명 이하로 분포하여 전처리 과정에서 대표성이 없다 판단하여 제외되었다.

8.3.1. <연구가설 1>

성별, 거주지역, 이민여부는 비경제활동여부에 영향을 미칠 것이다.

인구 정보가 비경제활동 여부에 영향을 미칠 것이라는 전제를 가지고 이 전제가 옳은지를 판단해보면서 분석을 진행해보도록 하겠다.

- 독립변수: 성별, 거주지역, 이민여구
- 종속변수: 직업 및 경제활동 상황- 비경제활동

여기서 구직을 원하지만 아직 일자리가 없는 실업자는 비경제활동인구에서 제외되었다. 구직활동을 하는 것이 곧 경제활동에 참여하고 있음을 의미하기 때문에 비경제활동 인구에는 포함되지 않았다. 비경제활동 인구에는 학생, 가정주부 등이 포함되어있으며 은퇴자의 경우 따로 분류되어 해당 범주에 포함되지 않았다.

일반 선형 혼합 모델(GLMM: Generalized linear mixed model)

먼저 일반 선형 모델을 통해 각 단일 독립변수(성별, 이민여부, 거주지역)이 종속변수(비경제활동 여부)에 미치는 영향을 알아보고자한다.

- 일반 선형 혼합 모델(GLMM)은 각 독립변수가 종속변수에 미치는 단일한 영향력은 물론 독립변수의 조합에 따른 분석 결과를 도출한다. 예를 들어 성별이 종속변수에 미치는 영향, 이민이 미치는 영향을 각각 살펴볼 수 도 있고 이민자 중에서도 남성인 사람과 이민자 중 여성인 사람의 차이도 살펴볼 수 있다. 이렇게 다수의 독립변수의 조합을 고려한 분석모델이 일반 선형 혼합 모델이다. 해당 모델은 단일한 독립변수의 영향력을 고려하였을때보다 좀 더 세부적인 영향력을 살펴볼 수 있다.

(1) 일반 선형 혼합 모델(GLMM) 구축

```
#model1_2 <- glmer(prof_autres ~ Sexe*immi+immi*region+region*Sexe +region +(1|ID), data = immi2018, family = binomial(link='logit'))
```

성별과 이민, 이민과 지역, 지역과 성별 간의 조합을 독립변수 값으로 입력하였다.

```
summary(model1_2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: prof_autres ~ Sexe * immi + immi * region + region * Sexe + region +
## (1 | ID)
## Data: immi2018
##
##      AIC      BIC    logLik deviance df.resid
##  4431.9   4499.9  -2204.9   4409.9     3552
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -0.9418 -0.6671 -0.6619  1.2676  2.2628
##
## Random effects:
## Groups Name      Variance Std.Dev.
## ID      (Intercept) 2.054e-11 4.532e-06
## Number of obs: 3563, groups: ID, 3563
##
## Fixed effects:
##
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.05684    0.13773   -7.673 1.68e-14 ***
## SexeHomme        -0.57632    0.19666   -2.931  0.00338 **
## immiNon-immigré    0.23160    0.15066    1.537  0.12423
## regionMar        0.93693    0.22618    4.142 3.44e-05 ***
## regionLyon       0.42938    0.30372    1.414  0.15744
## SexeHomme:immiNon-immigré  0.59196    0.20398    2.902  0.00371 **
## immiNon-immigré:regionMar -0.53818    0.23300   -2.310  0.02090 *
## immiNon-immigré:regionLyon -0.23292    0.31341   -0.743  0.45737
## SexeHomme:regionMar -0.06338    0.16941   -0.374  0.70830
## SexeHomme:regionLyon -0.04722    0.21019   -0.225  0.82225
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Warning in abbreviate(rn, minlength = 11): 아스키 문자가 아닌 것을 사용한 약어입
## 니다
```

```
##
## Correlation of Fixed Effects:
```

```
## Warning in abbreviate(rn, minlength = 6): 아스키 문자가 아닌 것을 사용한 약어입
## 니다
```

```
##              (Intr) SexHm immNn- regnMr rgnLyn SxH:N- imN-:M imN-:L SxHm:M
## SexeHomme      -0.533
## immiNn-mmgr    -0.883  0.429
## regionMar      -0.446  0.018  0.353
## regionLyon     -0.344  0.036  0.276  0.218
## SxHm:mmNn-     0.465 -0.873 -0.527  0.067  0.025
## immNn-mmG:M    0.384  0.074 -0.434 -0.860 -0.189 -0.061
## immNn-mmG:L    0.302  0.024 -0.342 -0.187 -0.884 -0.010  0.225
## SxHm:rgnMr     0.143 -0.269  0.041 -0.323 -0.067 -0.007 -0.027 -0.018
## SxHm:rgnLy     0.100 -0.187  0.051 -0.075 -0.272 -0.040 -0.017 -0.050  0.258
## optimizer (Nelder-Mead) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

p값이 0.05보다 낮으면 해당 변수가 종속변수에 유의적인 영향을 미치는 것으로 해석할 수 있다. 뒤에 붙은 * 표시가 많을수록 값이 작다는 의미이다.

- 유의적인 변수: 파리지주여부(**Intercept**에 해당), 성별, 마르세유 거주여부, 성별과 이민을 모두 고려한 혼합 변수, 마르세유 거주와 이민 여부를 모두 고려한 혼합 변수 인 것으로 나타났다.

(2) 분석 결과 해석

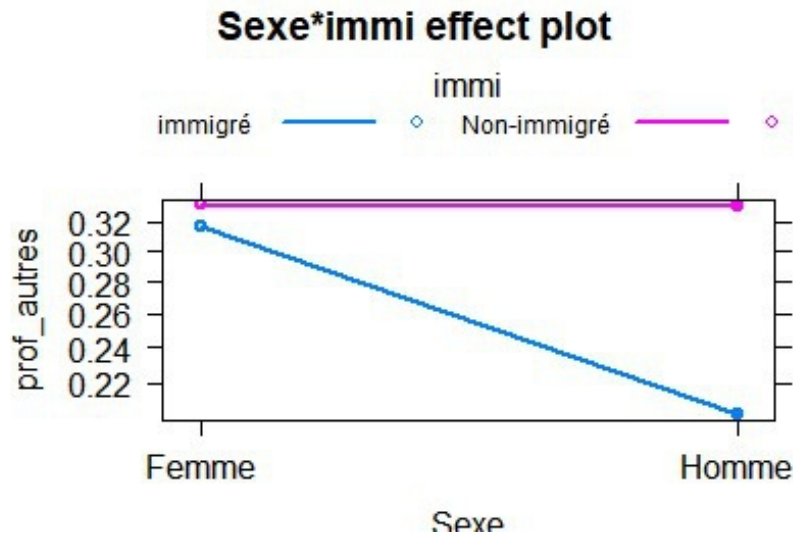
* 분석 결과 시각화

```
#plot(allEffects(model1_2), lines = list(multiline = TRUE)) # 중첩하여 그리기
```

- 다음은 해당 모델이 도출한 결론을 시각화한 것이다. 우리는 위에서 유의적인 변수로 도출된 변수의 영향을 중심으로 그래프를 살펴볼 것이다. 분석결과 p값이 0.05 이상으로 도출되어 유의적이지 않은 변수는 신뢰도가 낮다고 판단하여 그래프의 결과를 분석하지 않았다.

* 성별과 이민을 모두 고려한 혼합 변수

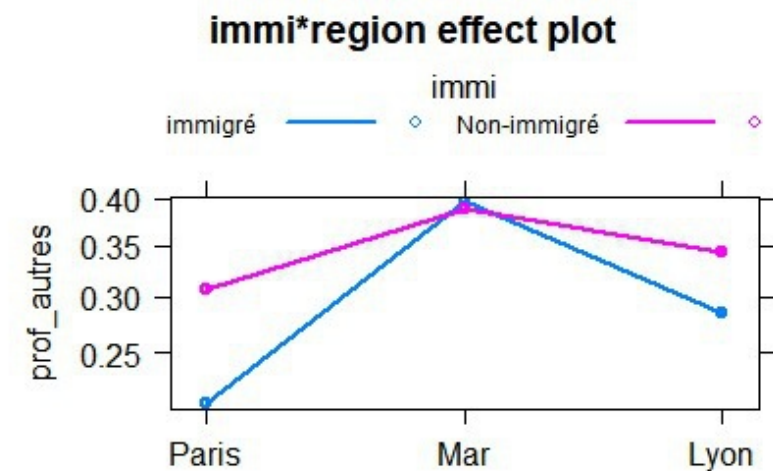
해당 분석에서 가장 중요한 내용은 이민의 영향이다. 우리는 성별과 이민의 영향을 모두 고려한 혼합 변수가 유의적인 것을 확인했는데 다음 그래프를 살펴보면 비이민자의 경우 성별의 영향이 거의 없는 것으로 나타나지만, 이민자의 경우 여성이 비경제활동 인구일 가능성이 더 높은 것으로 나타났다. 이를 통해 이민여부와 성별에 따라 비경제활동 인구일 가능성이 차이가 있는 것을 확인할 수 있었다.



<성별과 이민의 영향을 함께 고려한 비경제활동 여부 예측값 그래프>

* 마르세유 거주와 이민여부를 모두 고려한 혼합 변수

다음은 거주 지역과 이민여부를 모두 고려한 혼합 변수를 살펴볼 것이다. 마르세유의 경우 이민자와 비이민자의 비경제활동 여부 확률값이 거의 동일한 것을 확인할 수 있다.



<마르세유 거주와 이민의 영향을 함께 고려한 비경제활동 여부 예측값 그래프>

결론: 비경제활동 여부에서 비이민자의 경우 성별의 영향이 거의 없는 것으로 나타나지만 이민자의 경우 여성이 비경제활동 인구일 가능성이 더 높은 것으로 나타났다.

8.3.2. <연구가설 2>

성별, 거주지역, 이민여부는 **commercant** 직종 종사에 영향을 미칠 것이다.

- 독립변수: 성별, 거주지역, 이민여부
- 종속변수: 직업 및 경제활동 상황- **commercant(상업)** commercant(상업) 직업분류에는 수공업자, 상업종사자, 경영인이 속하였는데 이에는 도 소매업자나 자영업자 등이 포함되어있다.

일반선형혼합모델(GLMM)

(1) 일반선형혼합모델(GLMM) 구축

```
#model2_2 <- glmer(prof_commer ~ Sexe*immi+region +(1|ID), data = immi2018, family = binomial(link='logit'))

summary(model2_2)
```

```
## Cov prior : ID ~ wishart(df = 3.5, scale = Inf, posterior.scale = cov, common.scale = TRUE)
## Prior dev : -12.1084
##
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [bglmerMod]
## Family: binomial ( logit )
## Formula: prof_commer ~ Sexe * immi + region + (1 | ID)
## Data: immi2018
##
##      AIC      BIC   logLik deviance df.resid
##   429.6    472.8   -207.8    415.6     3556
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -0.001388 -0.001081 -0.000007  0.000000  0.190949
##
## Random effects:
##   Groups Name      Variance Std.Dev.
##   ID      (Intercept) 3204     56.61
## Number of obs: 3563, groups: ID, 3563
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -13.5214    20.5293  -0.659    0.510
## SexeHomme         0.3675    24.2729   0.015    0.988
## immiNon-immigré -10.1396    23.6751  -0.428    0.668
## regionMar        -9.2736    14.0390  -0.661    0.509
## regionLyon       -85.7508    17.3020 -4.956 7.19e-07 ***
## SexeHomme:immiNon-immigré  9.6368    27.9801   0.344    0.731
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr) SexHm immNn- regnMr rgnLyn
## SexeHomme    -0.835
## immiNn-mmgr -0.855  0.740
## regionMar   -0.086 -0.057 -0.068
## regionLyon  -0.017 -0.081 -0.128  0.196
## SxHm:mmNn-  0.724 -0.869 -0.844  0.046  0.090
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## unable to evaluate scaled gradient
## Model failed to converge: degenerate Hessian with 1 negative eigenvalues
```

확인결과 리옹 거주 여부만 유의적인 영향을 미치는 것으로 나타났으며 이민여부는 유의적인 영향을 미치지 않는 것으로 나타났다. 해당 분석의 목적은 직업 종사에 이민의 영향이 유의적인지 살펴보는 것이기 때문에 이민의 영향이 유의적이지 않은 연구가설은 세부적으로 결과를 해석하지 않았다.

8.3.3. <연구가설 3>

성별, 거주지역, 이민여부는 **Cadres(전문직)** 직종 종사에 영향을 미칠 것이다.

- 독립변수: 성별, 거주지역, 이민여부
- 종속변수: 직업 및 경제활동 상황- **Cadres(전문직)** 여기서 전문직 범주에는 고위직이나 교수와 기업의 임원이 포함되어있고, 변호사, 의사, 각종 기관의 임원, 영화감독 등이 포함되어있다.

일반 선형 혼합 모델(GLMM)

(1) 일반 선형 혼합 모델(GLMM) 생성

```
#model3_2 <- glmer(prof_Cadres ~ Sexe*immi+immi*region+region*Sexe +region+(1|ID), data=immi2018,family = binomial(
link='logit'))
summary(model3_2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: prof_Cadres ~ Sexe * immi + immi * region + region * Sexe + region +
## (1 | ID)
## Data: immi2018
##
##      AIC      BIC   logLik deviance df.resid
## 3420.4   3488.4 -1699.2   3398.4     3552
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -0.6633 -0.5687 -0.4669 -0.2720  4.5441
##
## Random effects:
## Groups Name             Variance Std.Dev.
## ID      (Intercept) 5.969e-11 7.726e-06
## Number of obs: 3563, groups: ID, 3563
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.52335    0.16188  -9.410 < 2e-16 ***
## SexeHomme           0.15771    0.22397   0.704 0.481328
## immiNon-immigr      0.39450    0.17665   2.233 0.025536 *
## regionMar          -1.50430    0.43323  -3.472 0.000516 ***
## regionLyon         -0.66567    0.41732  -1.595 0.110686
## SexeHomme:immiNon-immigr  0.14996    0.24088   0.623 0.533596
## immiNon-immigr :regionMar 0.02895    0.43200   0.067 0.946566
## immiNon-immigr :regionLyon 0.04085    0.41787   0.098 0.922131
## SexeHomme:regionMar    0.09091    0.27337   0.333 0.739475
## SexeHomme:regionLyon    0.16111    0.25542   0.631 0.528192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Warning in abbreviate(rn, minlength = 11): 아스키 문자가 아닌 것을 사용한 약어입
## 니다
```

```
##
## Correlation of Fixed Effects:
```

```
## Warning in abbreviate(rn, minlength = 6): 아스키 문자가 아닌 것을 사용한 약어입
## 니다
```

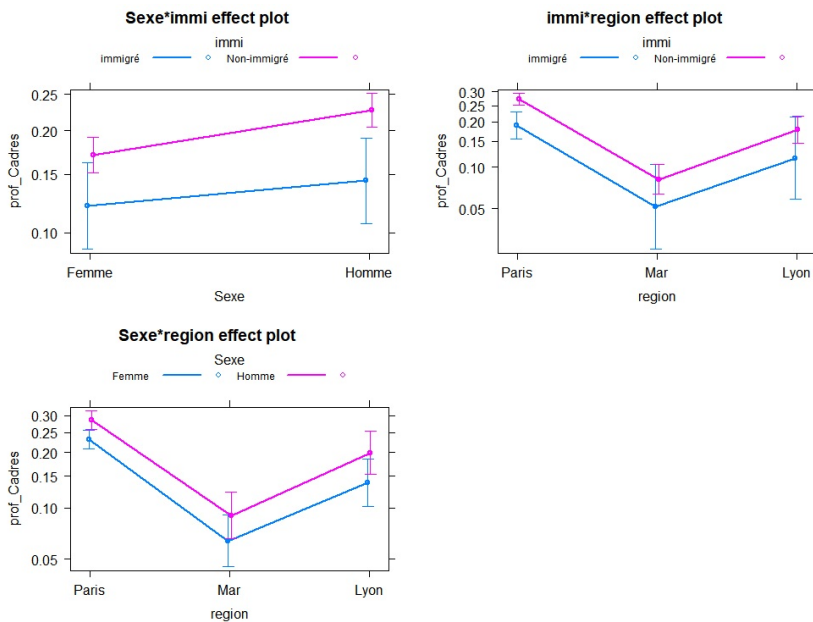
```
##              (Intr) SexHm immNn- regnMr rgnLyn SxH:N- imN-:M imN-:L SxHm:M
## SexeHomme      -0.662
## immiNn-mmgr   -0.902  0.585
## regionMar     -0.212  0.003  0.157
## regionLyon    -0.225  0.011  0.169  0.096
## SxHm:mmNn-    0.595 -0.898 -0.660  0.050  0.043
## immNn-mm:g:M  0.187  0.035 -0.207 -0.885 -0.082 -0.036
## immNn-mm:g:L  0.201  0.025 -0.222 -0.082 -0.894 -0.025  0.098
## SxHm:rgnMr    0.075 -0.112  0.040 -0.342 -0.043 -0.052 -0.002 -0.001
## SxHm:rgnLy    0.075 -0.113  0.049 -0.045 -0.318 -0.064  0.000 -0.011  0.143
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

- 유의적인 변수: 파리거주여부, 이민여부, 마르세유 거주 여부

(2) 결과 해석

```
#plot(allEffects(model3_2), lines = list(multiline = TRUE), confint = list(style = "auto")) # 중첩+오차표기
```

- 그래프 확인결과 전체적인 그래프에서 비이민자가 이민자보다 전문직 종사에 높은 확률을 나타내었다. 따라서 이민자일수록 전문직 종사 확률이 낮았던 것을 확인할 수 있었다.
- 또한 지역의 경우 파리가 마르세유보다 더 높은 확률을 나타내고 있었으며 리옹의 경우 오차 범위가 중첩되므로 유의적인 영향을 미친다고 해석하기 어려웠던 것으로 추측된다.
- 성별의 경우도 마찬가지로 여성과 남성의 오차범위가 상당부분 겹쳐져 나타나 성별이 달라지는 것으로 종사확률이 확연하게 달라진다고 해석하기 어려웠다.



전문직 종사 여부 예측값 그래프

8.3.4. <연구가설 4>

성별, 거주지역, 이민여부는 **professions intermédiaires** 직종 종사에 영향을 미칠 것이다.

- 독립변수: 성별, 거주지역, 이민여부
- 종속변수: 직업 및 경제활동 상황- **professions intermédiaires(중간직)** 여기서 중간직이란 임원과 직원 사이 직업군으로 학교교사, 중간관리직, 성직자, 경찰, 군인, 간호사와 같은 직업을 일컫는 프랑스에서 사용하는 직업 분류이다.

일반화 선형 혼합 모델(GLMM)

(1) 일반 선형 혼합 모델(GLMM) 생성

```
# model4_2 <- glmer(prof_Profess ~ Sexe*immi+region+(1|ID), data=immi2018, family = binomial(link='logit'))

summary(model4_2)
```



```
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: prof_Profess ~ Sexe * immi + immi * region + region * Sexe +
## region + (1 | ID)
## Data: immi2018
##
##           AIC          BIC    logLik deviance df.resid
##    1469.0     1537.0    -723.5   1447.0     3552
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -0.002714 -0.001756 -0.001486 -0.000022  0.107465
##
## Random effects:
## Groups Name             Variance Std.Dev.
## ID      (Intercept) 3950      62.85
## Number of obs: 3563, groups: ID, 3563
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -29.446     11.229  -2.622  0.00873 **
## SexeHomme        16.432     15.511   1.059  0.28942
## immiNon-immigré  17.657     12.154   1.453  0.14628
## regionMar      -11.187     23.246  -0.481  0.63034
## regionLyon      -8.392     25.508  -0.329  0.74214
## SexeHomme:immiNon-immigré -17.620     16.610  -1.061  0.28879
## immiNon-immigré:regionMar  4.647     23.464   0.198  0.84302
## immiNon-immigré:regionLyon  7.504     25.717   0.292  0.77044
## SexeHomme:regionMar   3.051     15.348   0.199  0.84244
## SexeHomme:regionLyon  -7.606     16.574  -0.459  0.64628
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr) SexHmM immNn- regnMr rgnLyn SxH:N- imN-:M imN-:L SxHm:M
## SexeHomme      -0.626
## immiNn-mmgr   -0.903  0.546
## regionMar     -0.310  0.026  0.245
## regionLyon    -0.282  0.022  0.222  0.152
## SxHmM:mmNn-   0.547 -0.874 -0.605  0.052  0.049
## immNn-mmM:M   0.279  0.019 -0.309 -0.903 -0.132 -0.032
## immNn-mmM:L   0.254  0.018 -0.281 -0.132 -0.907 -0.030  0.156
## SxHmM:rgnMr   0.105 -0.168  0.032 -0.323 -0.070 -0.082  0.047  0.011
## SxHmM:rgnLy   0.096 -0.154  0.031 -0.071 -0.320 -0.078  0.011  0.052  0.231
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## unable to evaluate scaled gradient
## Model failed to converge: degenerate Hessian with 3 negative eigenvalues
```

- 중간직 종사여부에 경우 파리 거주 여부 이외에 유의적인 영향을 미치는 변수가 없었다. 즉, 이민여부가 중간직 종사 여부에 유의적인 영향을 미치지 않는 것으로 확인되었다.

8.3.5. <연구가설 5>

성별, 거주지역, 이민여부는 **Employés** 직종 종사에 영향을 미칠 것이다.

- 독립변수: 성별, 거주지역, 이민여부
- 종속변수: 직업 및 경제활동 상황- **Employés** Employés에 속하는 직업으로는 공공서비스요원(공무원), 영업사원(판매원, 텔레마케터 등), 서비스업 종사자(미용사, 웨이터, 호텔직원) 등이 속한다.

일반화 선형 모델(GLMM)

(1) 일반화 선형 모델(GLMM) 생성

```
# model5_2 <- glmer(prof_Employes ~ Sexe*immi+immi*region+region*Sexe +region +(1|ID),data=immi2018, family = binomial(link='logit'))
summary(model5_2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: prof_Employes ~ Sexe * immi + immi * region + region * Sexe +
## region + (1 | ID)
## Data: immi2018
##
##           AIC          BIC    logLik deviance df.resid
##    2453.7    2521.6   -1215.8    2431.7     3552
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -0.5703 -0.3842 -0.3265 -0.2655  3.7660
##
## Random effects:
## Groups Name             Variance Std.Dev.
## ID      (Intercept) 6.432e-13 8.02e-07
## Number of obs: 3563, groups: ID, 3563
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.12331    0.14416  -7.792 6.59e-15 ***
## SexeHomme         -0.78962    0.22888  -3.450 0.000561 ***
## immiNon-immigré   -1.11530    0.17618  -6.330 2.45e-10 ***
## regionMar         -0.13860    0.28340  -0.489 0.624805
## regionLyon        -0.03305    0.35972  -0.092 0.926792
## SexeHomme:immiNon-immigré  0.37621    0.25964    1.449 0.147343
## immiNon-immigré:regionMar  0.73031    0.30559    2.390 0.016856 *
## immiNon-immigré:regionLyon 0.44537    0.38811    1.148 0.251162
## SexeHomme:regionMar  -0.25177    0.26215  -0.960 0.336847
## SexeHomme:regionLyon  -0.27055    0.33022  -0.819 0.412608
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr) SexHm immNn- regnMr rgnLyn SxH:N- imN-:M imN-:L SxHm:M
## SexeHomme      -0.533
## immiNn-mmgr   -0.782  0.368
## regionMar     -0.410  0.086  0.272
## regionLyon    -0.329  0.078  0.221  0.171
## SxHm:mmNn-    0.407 -0.764 -0.520  0.035  0.014
## immNn-mmG:M   0.339 -0.003 -0.434 -0.835 -0.132  0.025
## immNn-mmG:L   0.273 -0.012 -0.348 -0.131 -0.844  0.033  0.192
## SxHm:rgnMr    0.130 -0.244  0.107 -0.298 -0.080 -0.157 -0.036 -0.017
## SxHm:rgnLy    0.104 -0.195  0.084 -0.082 -0.277 -0.122 -0.015 -0.042  0.271
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

- 유의적인 변수: 파리지주여부, 성별, 이민여부, 마르세유 거부와 이민의 영향을 모두 고려한 혼합 변수

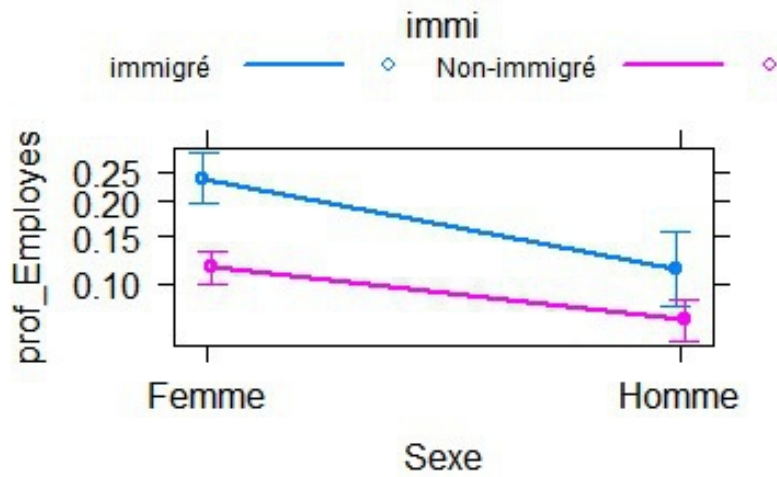
(2) 결과 해석

- 해당 분석에서는 성별과 이민의 영향을 중점적으로 살펴보았다.

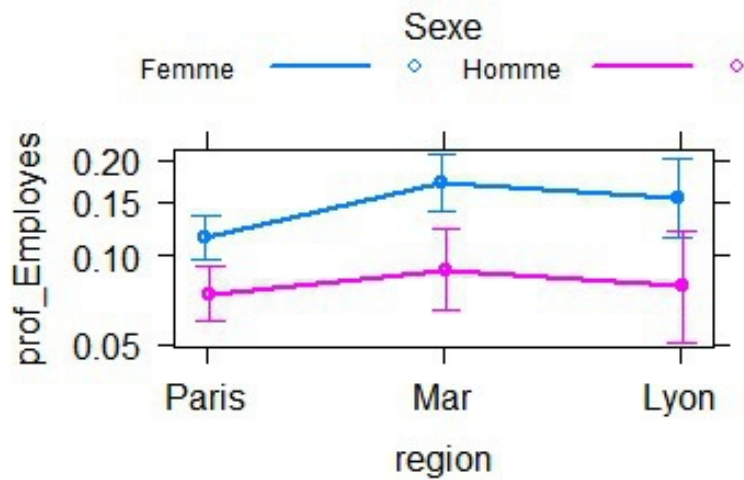
```
#plot(allEffects(model5_2),lines = list(multiline = TRUE),confint = list(style = "auto"))
```

- 성별 다음 그래프를 살펴보면 여성일수록 해당 직종에 종사할 확률이 모든 그래프에서 높은 것을 확인할 수 있다.

Sexe*immi effect plot



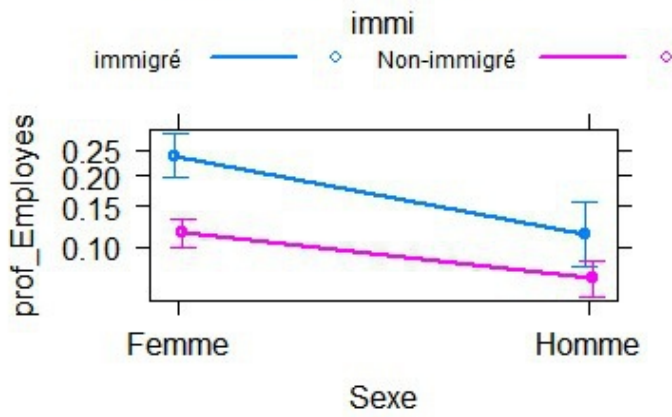
Sexe*region effect plot



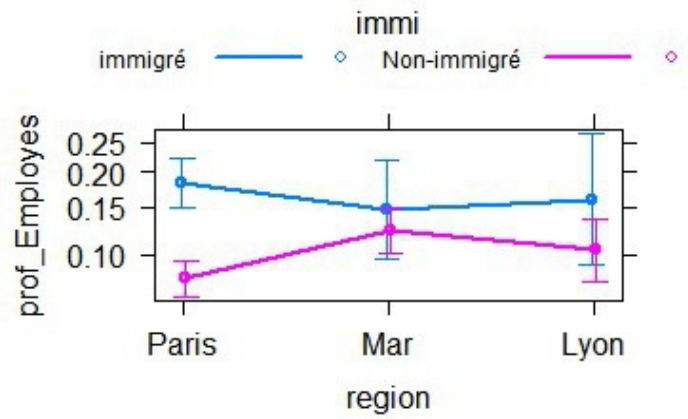
직원 종사여부에 관한 성별의 영향

- 이민 다음의 그래프를 살펴보았을 때 이민자일수록 해당 직종의 종사할 확률이 모든 그래프에서 높은 것을 확인할 수 있었다.

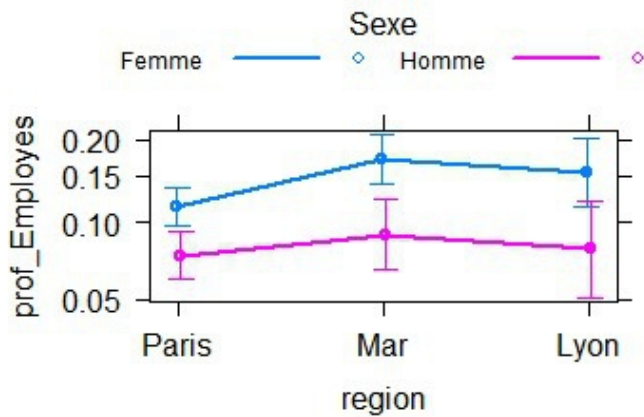
Sexe*immi effect plot



immi*region effect plot



Sexe*region effect plot



직원 종사여부에 관한 이민의 영향

- 결과: 따라서 성별이 여성일수록, 이민자일수록 해당 직종에 종사할 확률이 높아졌다.

8.3.6. <연구가설 6>

성별, 거주지역, 이민여부는 **Ouvriers** 직종 종사에 영향을 미칠 것이다.

- 독립변수: 성별, 거주지역, 이민 여부
- 종속변수: 직업 및 경제활동 상황- **Ouvriers(단순노동직)** Ouvriers(단순노동직)에는 산업숙련노동자(건설업 현장 작업자, 전기기술자), 자동차 정비사 등이 건설업업 정비업 등에 종사하는 직업이 여기에 속해있다.

일반 선형 혼합 모델(GLMM)

(1) 일반 선형 혼합 모델(GLMM) 모델 생성

```
#model6_2 <- glmer(prof_Ouvriers ~ Sexe*immi +region +(1|ID),data=immi2018,family = binomial(link='logit'))
summary(model6_2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: prof_Ouvriers ~ Sexe * immi + immi * region + region * Sexe +
## region + (1 | ID)
## Data: immi2018
##
##          AIC          BIC    logLik deviance df.resid
##    566.0      633.9   -272.0     544.0     3552
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -0.002689 -0.001162 -0.000512 -0.000166  0.077671
##
## Random effects:
## Groups Name             Variance Std.Dev.
## ID      (Intercept) 3890      62.37
## Number of obs: 3563, groups: ID, 3563
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.344e+01  5.325e-04  -25244   <2e-16 ***
## SexeHomme          -1.709e+00  5.325e-04   -3209   <2e-16 ***
## immiNon-immigr     -6.760e-02  5.325e-04    -127   <2e-16 ***
## regionMar           9.039e-01  5.325e-04    1697   <2e-16 ***
## regionLyon        -1.017e+00  5.325e-04   -1910   <2e-16 ***
## SexeHomme:immiNon-immigr  -2.189e+00  5.325e-04   -4110   <2e-16 ***
## immiNon-immigr :regionMar -3.460e+00  5.325e-04   -6498   <2e-16 ***
## immiNon-immigr :regionLyon 1.302e+00  5.325e-04    2445   <2e-16 ***
## SexeHomme:regionMar  1.746e+00  5.325e-04    3279   <2e-16 ***
## SexeHomme:regionLyon  4.360e+00  5.325e-04    8187   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Warning in abbreviate(rn, minlength = 11): 아스키 문자가 아닌 것을 사용한 약어입
## 니다
```

```
##
## Correlation of Fixed Effects:
```

```
## Warning in abbreviate(rn, minlength = 6): 아스키 문자가 아닌 것을 사용한 약어입
## 니다
```

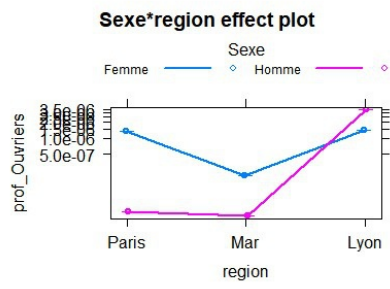
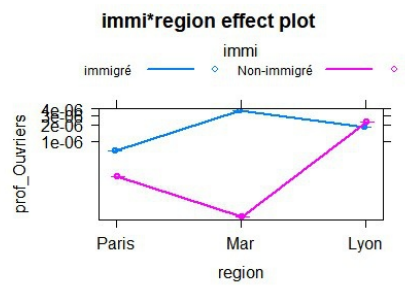
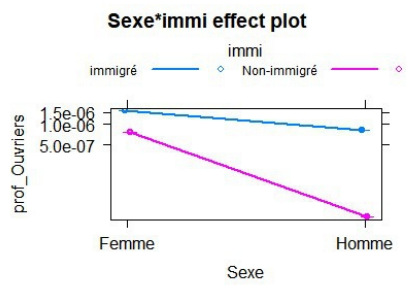
```
##              (Intr) SexHm immNn- regnMr rgnLyn SxH:N- imN-:M imN-:L SxHm:M
## SexeHomme      0.000
## immiNn-mmgr  0.000  0.000
## regionMar    0.000  0.000  0.000
## regionLyon   0.000  0.000  0.000  0.000
## SxHm:mmNn-   0.000  0.000  0.000  0.000  0.000
## immNn-mm:g:M 0.000  0.000  0.000  0.000  0.000  0.000
## immNn-mm:g:L 0.000  0.000  0.000  0.000  0.000  0.000  0.000
## SxHm:rgnMr   0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000
## SxHm:rgnLy   0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.134059 (tol = 0.002, component 1)
## Model is nearly unidentifiable: very large eigenvalue
## - Rescale variables?
```

- 결과: 모든 변인이 유의적인 것으로 나타났다.

(2) 결과해석

```
#plot(allEffects(model6_2),lines = list(multiline = TRUE),confint = list(style = "auto"))
```

- 모든 변인이 유의적이었기 때문에 우리는 이민여부만 중점적으로 살펴보고자 한다.
- 이민
- 시각화 결과를 살펴보면 이민자일수록 해당 직군에 종사할 확률이 높았다. 특이한 점이 있다면 리옹지역의 결과였는데 해당 지역은 비이민자의 확률이 근소하게 높은 것을 확인할 수 있었다.
- 또한 성별과 이민을 함께 고려한 그래프를 살펴보면 이민자의 경우 성별에 따라 종사 확률의 차이 기울기가 완만했지만 비이민자의 경우 성별에 따른 종사 확률이 가파른 것으로 나타났다. 따라서 이민자의 경우 성별이 단순노동직 종사 확률에 미치는 영향이 비이민자보다 상대적으로 낮다고 결론지을 수 있었다.



단순노동직 종사 여부에 관한 이민의 영향