

Group_project_5014

Xinyi Song

11/21/2020

Social Disparity - Income

```
# 2019 region
library(plotrix)
library(treemap)
library(reshape2)
library(ggplot2)
Allraces<- read.csv("~/Desktop/VTCourse/STAT 5014/Group Project/Total work experience-both sexes/2019/A
Allraces = as.data.frame(Allraces[-1,])
colnames(Allraces) = c('Region','Total', 'Total with Income', '$2499', '$2500-$4999', '$5000-$7499', '$
'$10000-$12499', '$12500-$14999', '$15000-$17499', '$17500-$19999', '$20000-$224
'$22500-$24999', '$25000-$27499', '$27500-$29999', '$30000-$32499', '$32500-$374
'$42500-$44999', '$45000-$47499', '$47500-$49999', '$50000-$52499', '$52500-$5499
'$100000-'
)
all_race_income_range = Allraces[,4:44]
colnames(all_race_income_range) = c('$2499', '$2500-$4999', '$5000-$7499', '$7500-$9999',
'$10000-$12499', '$12500-$14999', '$15000-$17499', '$17500-$19999', '$20000-$224
'$22500-$24999', '$25000-$27499', '$27500-$29999', '$30000-$32499', '$32500-$374
'$42500-$44999', '$45000-$47499', '$47999-$49999', '$50000-$52499', '$52500-$5499
'$100000-')
# Change the range to $10000
range_combine_allr = matrix(0, 4, 9)
for (i in 1:dim(all_race_income_range)[1]){
  for (j in 1:dim(all_race_income_range)[2]){
all_race_income_range[i,j] = as.numeric(gsub(",", "", all_race_income_range[i,j]))
  }
}
for (i in 1:dim(all_race_income_range)[1]){
  range_combine_allr[i,1] = sum(as.numeric(all_race_income_range[i,1:5])) # 0 - 12499
  range_combine_allr[i,2] = sum(as.numeric(all_race_income_range[i,6:10])) #12500 - 24999
  range_combine_allr[i,3] = sum(as.numeric(all_race_income_range[i,11:15])) # 25000 - 37499
  range_combine_allr[i,4] = sum(as.numeric(all_race_income_range[i,16:20])) # 37500 - 49999
  range_combine_allr[i,5] = sum(as.numeric(all_race_income_range[i,21:25])) # 50000 - 62499
  range_combine_allr[i,6] = sum(as.numeric(all_race_income_range[i,26:30])) # 62500 - 74999
  range_combine_allr[i,7] = sum(as.numeric(all_race_income_range[i,31:35])) # 75000- 87499
  range_combine_allr[i,8] = sum(as.numeric(all_race_income_range[i,36:40])) # 87500 - 99999
  range_combine_allr[i,9] = sum(as.numeric(all_race_income_range[i,41])) # >= 100000
}
range_combine_allr = as.data.frame(range_combine_allr)
rownames(range_combine_allr) = c('Northeast', 'Midwest', 'South', 'West')
```

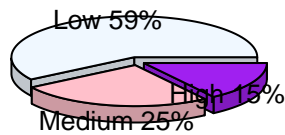
```

#colnames(range_combine_allr) = c('0to12499', '12500to24999', '25000to37499', '37500to49999', '50000to62
colnames(range_combine_allr) = seq(0,100000, 12500)
# After divide the total income range data, I tried to divide it into low, medium and high level
income_level_all = matrix(0, 4, 3)
for (i in 1:4){
  income_level_all[i,1] = sum(range_combine_allr[i,1:4]) # 0 - 49999 Low Level
  income_level_all[i,2] = sum(range_combine_allr[i,5:8]) # 49999- 99999 Medium Level
  income_level_all[i,3] = sum(range_combine_allr[i,9]) # >= 199999 High Level
}
income_level_all = as.data.frame(income_level_all)
colnames(income_level_all) = c('Low', 'Medium', 'High')
rownames(income_level_all) = c('Northeast', 'Midwest', 'South', 'West')

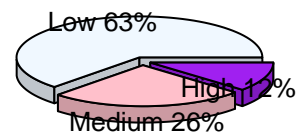
# Pie plot
###
par(mfrow=c(2,2))
x <- as.vector(unlist(income_level_all[1,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.1,labelcex =0.8, radius = 1.5,
      main="Income Level of Northeast in 2019", col = c('aliceblue', 'pink', 'purple'))
x <- as.vector(unlist(income_level_all[2,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.1, labelcex = 0.8, radius = 1.5,
      main="Income Level of Midwest in 2019", cex= 0.5, col = c('aliceblue', 'pink', 'purple'))
x <- as.vector(unlist(income_level_all[3,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.05,labelcex = 0.8, radius = 1.5,
      main="Income Level of South in 2019", col = c('aliceblue', 'pink', 'purple'))
x <- as.vector(unlist(income_level_all[4,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.05, labelcex = 0.8, radius = 1.5,
      main="Income Level of West in 2019", col = c('aliceblue', 'pink', 'purple'))

```

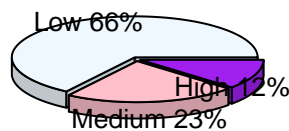
Income Level of Northeast in 2019



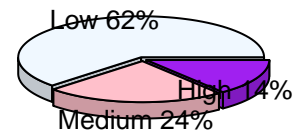
Income Level of Midwest in 2019



Income Level of South in 2019



Income Level of West in 2019



```
# 2018 region
allraces_2018 <- read.csv("~/Desktop/VTCourse/STAT 5014/Group Project/Total work experience-both sexes/2018/allraces_2018.csv")
allraces_2018 = as.data.frame(allraces_2018 [-1,])
colnames(allraces_2018) = c('Region', 'Total', 'Total with Income', '$2499', '$2500-$4999', '$5000-$7499', '$7500-$9999', '$10000-$12499', '$12500-$14999', '$15000-$17499', '$17500-$19999', '$20000-$22499', '$22500-$24999', '$25000-$27499', '$27500-$29999', '$30000-$32499', '$32500-$34999', '$35000-$37499', '$37500-$39999', '$40000-$42499', '$42500-$44999', '$45000-$47499', '$47500-$49999', '$50000-$52499', '$52500-$54999', '$55000-$57499', '$57500-$59999', '$60000-$62499', '$62500-$64999', '$65000-$67499', '$67500-$69999', '$70000-$72499', '$72500-$74999', '$75000-$77499', '$77500-$79999', '$80000-$82499', '$82500-$84999', '$85000-$87499', '$87500-$89999', '$90000-$92499', '$92500-$94999', '$95000-$97499', '$97500-$99999', '$100000-')
)

all_race_income_range_2018 = allraces_2018[,4:44]
colnames(all_race_income_range_2018) = c('$2499', '$2500-$4999', '$5000-$7499', '$7500-$9999', '$10000-$12499', '$12500-$14999', '$15000-$17499', '$17500-$19999', '$20000-$22499', '$22500-$24999', '$25000-$27499', '$27500-$29999', '$30000-$32499', '$32500-$34999', '$35000-$37499', '$37500-$39999', '$40000-$42499', '$42500-$44999', '$45000-$47499', '$47500-$49999', '$50000-$52499', '$52500-$54999', '$55000-$57499', '$57500-$59999', '$60000-$62499', '$62500-$64999', '$65000-$67499', '$67500-$69999', '$70000-$72499', '$72500-$74999', '$75000-$77499', '$77500-$79999', '$80000-$82499', '$82500-$84999', '$85000-$87499', '$87500-$89999', '$90000-$92499', '$92500-$94999', '$95000-$97499', '$97500-$99999', '$100000-')

range_combine_allr_2018 = matrix(0, 4, 9)
for (i in 1:dim(all_race_income_range_2018)[1]){
  for (j in 1:dim(all_race_income_range_2018)[2]){
    all_race_income_range_2018[i,j] = as.numeric(gsub("-", "", all_race_income_range_2018[i,j]))
  }
}

for (i in 1:dim(all_race_income_range_2018)[1]){
  range_combine_allr_2018[i,1] = sum(as.numeric(all_race_income_range_2018[i,1:5])) # 0 - 12499
  range_combine_allr_2018[i,2] = sum(as.numeric(all_race_income_range_2018[i,6:10])) # 12500 - 24999
  range_combine_allr_2018[i,3] = sum(as.numeric(all_race_income_range_2018[i,11:15])) # 25000 - 37499
  range_combine_allr_2018[i,4] = sum(as.numeric(all_race_income_range_2018[i,16:20])) # 37500 - 49999
  range_combine_allr_2018[i,5] = sum(as.numeric(all_race_income_range_2018[i,21:25])) # 50000 - 62499
  range_combine_allr_2018[i,6] = sum(as.numeric(all_race_income_range_2018[i,26:30])) # 62500 - 74999
  range_combine_allr_2018[i,7] = sum(as.numeric(all_race_income_range_2018[i,31:35])) # 75000 - 87499
  range_combine_allr_2018[i,8] = sum(as.numeric(all_race_income_range_2018[i,36:40])) # 87500 - 99999
  range_combine_allr_2018[i,9] = sum(as.numeric(all_race_income_range_2018[i,41])) # >= 100000
}
```

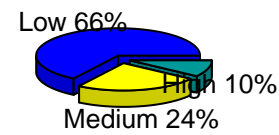
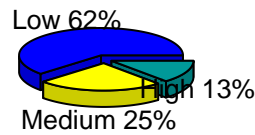
```

range_combine_allr_2018 = as.data.frame(range_combine_allr_2018)
rownames(range_combine_allr_2018) = c('Northeast', 'Midwest', 'South', 'West')
#colnames(range_combine_allr) = c('0to12499', '12500to24999', '25000to37499', '37500to49999', '50000to62
colnames(range_combine_allr_2018) = seq(0,100000, 12500)
# After divide the total income range data, I tried to divide it into low, medium and high level
income_level_all_2018 = matrix(0, 4, 3)
for (i in 1:4){
  income_level_all_2018[i,1] = sum(range_combine_allr_2018[i,1:4]) # 0 - 49999 Low Level
  income_level_all_2018[i,2] = sum(range_combine_allr_2018[i,5:8]) # 49999- 99999 Medium Level
  income_level_all_2018[i,3] = sum(range_combine_allr_2018[i,9]) # >= 99999 High Level
}
income_level_all_2018 = as.data.frame(income_level_all_2018)
colnames(income_level_all_2018) = c('Low', 'Medium', 'High')
rownames(income_level_all_2018) = c('Northeast', 'Midwest', 'South', 'West')

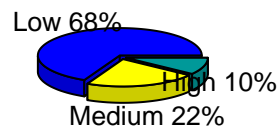
# Pie plot
###
par(mfrow=c(2,2))
x <- as.vector(unlist(income_level_all_2018[1,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.1,labelcex = 0.8,
      main="Income Level Percent of Northeast of 2018", col = c('blue', 'yellow', '#009999'))
x <- as.vector(unlist(income_level_all_2018[2,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.1, labelcex = 0.8,col = c('blue', 'yellow', '#009999'),
      main="Income Level Percent of Midwest of 2018")
x <- as.vector(unlist(income_level_all_2018[3,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.05, labelcex = 0.8,col = c('blue', 'yellow', '#009999'),
      main="Income Level Percent of South of 2018")
x <- as.vector(unlist(income_level_all_2018[4,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.05, labelcex = 0.8,col = c('blue', 'yellow', '#009999'),
      main="Income Level Percent of West of 2018")

```

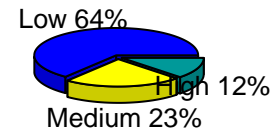
Income Level Percent of Northeast of 2018



Income Level Percent of South of 2018



Income Level Percent of West of 2018



```
# 2017 region
allraces_2017 <- read.csv("~/Desktop/VTCourse/STAT 5014/Group Project/Total work experience-both sexes/")
allraces_2017 = as.data.frame(allraces_2017[-1,])
colnames(allraces_2017) = c('Region','Total', 'Total with Income', '$2499', '$2500-$4999', '$5000-$7499',
'$10000-$12499', '$12500-$14999', '$15000-$17499', '$17500-$19999', '$20000-$22499',
'$22500-$24999', '$25000-$27499', '$27500-$29999', '$30000-$32499', '$32500-$37499',
'$42500-$44999', '$45000-$47499', '$47500-$49999', '$50000-$52499', '$52500-$54999',
'$100000-'
)
all_race_income_range_2017 = allraces_2017[,4:44]
colnames(all_race_income_range_2017) = c('$2499', '$2500-$4999', '$5000-$7499', '$7500-$9999',
'$10000-$12499', '$12500-$14999', '$15000-$17499', '$17500-$19999', '$20000-$22499',
'$22500-$24999', '$25000-$27499', '$27500-$29999', '$30000-$32499', '$32500-$37499',
'$42500-$44999', '$45000-$47499', '$47999-$49999', '$50000-$52499', '$52500-$54999',
'$100000-')
# Change the range to $10000
range_combine_allr_2017 = matrix(0, 4, 9)
for (i in 1:dim(all_race_income_range_2017)[1]){
  for (j in 1:dim(all_race_income_range_2017)[2]){
    all_race_income_range_2017[i,j] = as.numeric(gsub("-", "", all_race_income_range_2017[i,j]))
  }
}
for (i in 1:dim(all_race_income_range_2017)[1]){
  range_combine_allr_2017[i,1] = sum(as.numeric(all_race_income_range_2017[i,1:5])) # 0 - 12499
  range_combine_allr_2017[i,2] = sum(as.numeric(all_race_income_range_2017[i,6:10])) # 12500 - 24999
  range_combine_allr_2017[i,3] = sum(as.numeric(all_race_income_range_2017[i,11:15])) # 25000 - 37499
  range_combine_allr_2017[i,4] = sum(as.numeric(all_race_income_range_2017[i,16:20])) # 37500 - 49999
  range_combine_allr_2017[i,5] = sum(as.numeric(all_race_income_range_2017[i,21:25])) # 50000 - 62499
  range_combine_allr_2017[i,6] = sum(as.numeric(all_race_income_range_2017[i,26:30])) # 62500 - 74999
  range_combine_allr_2017[i,7] = sum(as.numeric(all_race_income_range_2017[i,31:35])) # 75000 - 87499
  range_combine_allr_2017[i,8] = sum(as.numeric(all_race_income_range_2017[i,36:40])) # 87500 - 99999
  range_combine_allr_2017[i,9] = sum(as.numeric(all_race_income_range_2017[i,41])) # >= 100000
}
```

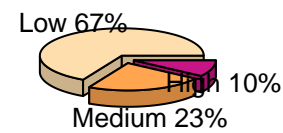
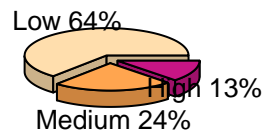
```

}
range_combine_allr_2017 = as.data.frame(range_combine_allr_2017)
rownames(range_combine_allr_2017) = c('Northeast', 'Midwest', 'South', 'West')
#colnames(range_combine_allr) = c('0to12499', '12500to24999', '25000to37499', '37500to49999', '50000to62
colnames(range_combine_allr_2017) = seq(0,100000, 12500)
# After divide the total income range data, I tried to divide it into low, medium and high level
income_level_all_2017 = matrix(0, 4, 3)
for (i in 1:4){
  income_level_all_2017[i,1] = sum(range_combine_allr_2017[i,1:4]) # 0 - 49999 Low Level
  income_level_all_2017[i,2] = sum(range_combine_allr_2017[i,5:8]) # 49999- 99999 Medium Level
  income_level_all_2017[i,3] = sum(range_combine_allr_2017[i,9]) # >= 199999 High Level
}
income_level_all_2017 = as.data.frame(income_level_all_2017)
colnames(income_level_all_2017) = c('Low', 'Medium', 'High')
rownames(income_level_all_2017) = c('Northeast', 'Midwest', 'South', 'West')

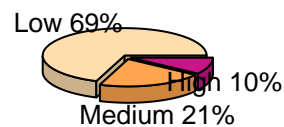
par(mfrow=c(2,2))
x <- as.vector(unlist(income_level_all_2017[1,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.1,labelcex = 0.8,
      main="Income Level Percent of Northeast of 2017", col = c('navajowhite1', 'tan1', 'mediumvioletred'))
x <- as.vector(unlist(income_level_all_2017[2,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.1, labelcex = 0.8,col = c('navajowhite1', 'tan1', 'mediumvioletred'),
      main="Income Level Percent of Midwest of 2017")
x <- as.vector(unlist(income_level_all_2017[3,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.05, labelcex = 0.8,col = c('navajowhite1', 'tan1', 'mediumvioletred'),
      main="Income Level Percent of South of 2017")
x <- as.vector(unlist(income_level_all_2017[4,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.05, labelcex = 0.8,col = c('navajowhite1', 'tan1', 'mediumvioletred'),
      main="Income Level Percent of West of 2017")

```

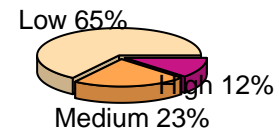
Income Level Percent of Northeast of 2017



Income Level Percent of South of 2017



Income Level Percent of West of 2017



```
# Social disparity about Education level
Education_2017 <- read.csv("~/Desktop/VTCourse/STAT 5014/Group Project/Total work experience-both sexes,
Education_2017 <- Education_2017[-c(1,2,8),]
Education_2017[,1] <- c('9th_Grade', '12th_Grade', 'High_School', 'College', 'Associate', 'Bachelor', 'M
rowname_edu = Education_2017[,1]
Education_2017 <- as.data.frame(Education_2017[,-(1:3)])
rownames(Education_2017) <- rowname_edu
colnames(Education_2017) = c('$2499', '$2500-$4999', '$5000-$7499', '$7500-$9999',
'$10000-$12499', '$12500-$14999', '$15000-$17499', '$17500-$19999', '$20000-$22499',
'$22500-$24999', '$25000-$27499', '$27500-$29999', '$30000-$32499', '$32500-$37499',
'$42500-$44999', '$45000-$47499', '$47999-$49999', '$50000-$52499', '$52500-$54999',
'$100000-')
edu_income_range_2017 = matrix(0, 9, 9)
for (i in 1:dim(Education_2017)[1]){
  for (j in 1:dim(Education_2017)[2]){
    Education_2017[i,j] = as.numeric(gsub("-", "", Education_2017[i,j]))
  }
}
for (i in 1:dim(edu_income_range_2017)[1]){
  edu_income_range_2017[i,1] = sum(as.numeric(Education_2017[i,1:5])) # 0 - 12499
  edu_income_range_2017[i,2] = sum(as.numeric(Education_2017[i,6:10])) # 12500 - 24999
  edu_income_range_2017[i,3] = sum(as.numeric(Education_2017[i,11:15])) # 25000 - 37499
  edu_income_range_2017[i,4] = sum(as.numeric(Education_2017[i,16:20])) # 37500 - 49999
  edu_income_range_2017[i,5] = sum(as.numeric(Education_2017[i,21:25])) # 50000 - 62499
  edu_income_range_2017[i,6] = sum(as.numeric(Education_2017[i,26:30])) # 62500 - 74999
  edu_income_range_2017[i,7] = sum(as.numeric(Education_2017[i,31:35])) # 75000 - 87499
  edu_income_range_2017[i,8] = sum(as.numeric(Education_2017[i,36:40])) # 87500 - 99999
  edu_income_range_2017[i,9] = sum(as.numeric(Education_2017[i,41])) # >= 100000
}
edu_income_range_2017 = as.data.frame(edu_income_range_2017)
rownames(edu_income_range_2017) = c('9th_Grade', '12th_Grade', 'High_School', 'College', 'Associate', 'B
#colnames(range_combine_allr) = c('0to12499', '12500to24999', '25000to37499', '37500to49999', '50000to62499',
```



```

colnames(edu_income_range_2017) = seq(0,100000, 12500)
# After divide the total income range data, I tried to divide it into low, medium and high level
income_edu_2017 = matrix(0, 9, 3)
for (i in 1:9){
  income_edu_2017[i,1] = sum(edu_income_range_2017[i,1:4]) # 0 - 49999 Low Level
  income_edu_2017[i,2] = sum(edu_income_range_2017[i,5:8]) # 49999- 99999 Medium Level
  income_edu_2017[i,3] = sum(edu_income_range_2017[i,9]) # >= 199999 High Level
}
income_edu_2017= as.data.frame(income_edu_2017)
colnames(income_edu_2017) = c('Low', 'Medium', 'High')
rownames(income_edu_2017) =c('9th_Grade', '12th_Grade', 'High_School', 'College', 'Associate','Bachelor')
income_edu_2017_mod = matrix(0, 3, 3)
for (i in 1:dim(income_edu_2017_mod)[2]){
  income_edu_2017_mod[1,i] = sum(income_edu_2017[1:3,i])
  income_edu_2017_mod[2,i] = sum(income_edu_2017[4:5,i])
  income_edu_2017_mod[3,i] = sum(income_edu_2017[6:9,i])
}
income_edu_2017_mod = as.data.frame(income_edu_2017_mod)
colnames(income_edu_2017_mod) = c('Low', 'Medium', "High")
rownames(income_edu_2017_mod) = c('HighSchool', 'College', 'Bachelor Above')

# Education 2018
Education_2018 <- read.csv("~/Desktop/VTCourse/STAT 5014/Group Project/Total work experience-both sexes.csv")
Education_2018 <- Education_2018[-c(1,2,8),]
Education_2018[,1] <- c('9th_Grade', '12th_Grade', 'High_School', 'College', 'Associate', 'Bachelor', 'Master')
rowname_edu = Education_2018[,1]
Education_2018 <- as.data.frame(Education_2018[,-(1:3)])
rownames(Education_2018) <- rowname_edu
colnames(Education_2018) = c('$2499', '$2500-$4999', '$5000-$7499', '$7500-$9999',
                             '$10000-$12499', '$12500-$14999', '$15000-$17499', '$17500-$19999', '$20000-$22499',
                             '$22500-$24999', '$25000-$27499', '$27500-$29999', '$30000-$32499', '$32500-$34999',
                             '$35000-$37499', '$37500-$39999', '$40000-$42499', '$42500-$44999', '$45000-$47499', '$47999-$49999', '$50000-$52499', '$52500-$54999',
                             '$55000-$57499', '$57500-$59999', '$60000-$62499', '$62500-$64999', '$65000-$67499', '$67500-$69999', '$70000-$72499', '$72500-$74999',
                             '$75000-$77499', '$77500-$79999', '$80000-$82499', '$82500-$84999', '$85000-$87499', '$87500-$89999', '$90000-$92499', '$92500-$94999',
                             '$95000-$97499', '$97500-$99999', '$100000-')
edu_income_range_2018 = matrix(0, 9, 9)
for (i in 1:dim(Education_2018)[1]){
  for (j in 1:dim(Education_2018)[2]){
    Education_2018[i,j] = as.numeric(gsub("-", "", Education_2018[i,j]))
  }
}
for (i in 1:dim(edu_income_range_2018)[1]){
  edu_income_range_2018[i,1] = sum(as.numeric(Education_2018[i,1:5])) # 0 - 12499
  edu_income_range_2018[i,2] = sum(as.numeric(Education_2018[i,6:10])) #12500 - 24999
  edu_income_range_2018[i,3] = sum(as.numeric(Education_2018[i,11:15])) # 25000 - 37499
  edu_income_range_2018[i,4] = sum(as.numeric(Education_2018[i,16:20])) # 37500 - 49999
  edu_income_range_2018[i,5] = sum(as.numeric(Education_2018[i,21:25])) # 50000 - 62499
  edu_income_range_2018[i,6] = sum(as.numeric(Education_2018[i,26:30])) # 62500 - 74999
  edu_income_range_2018[i,7] = sum(as.numeric(Education_2018[i,31:35])) # 75000- 87499
  edu_income_range_2018[i,8] = sum(as.numeric(Education_2018[i,36:40])) # 87500 - 99999
  edu_income_range_2018[i,9] = sum(as.numeric(Education_2018[i,41])) # >= 100000
}
edu_income_range_2018 = as.data.frame(edu_income_range_2018)
rownames(edu_income_range_2018) = c('9th_Grade', '12th_Grade', 'High_School', 'College', 'Associate', 'Bachelor')
#colnames(range_combine_allr) = c('0to12499', '12500to24999', '25000to37499', '37500to49999', '50000to62499', '62500to74999', '75000to87499', '87500to99999', '100000to125000')
colnames(edu_income_range_2018) = seq(0,100000, 12500)

```



```

# After divide the total income range data, I tried to divide it into low, medium and high level
income_edu_2018 = matrix(0, 9, 3)
for (i in 1:9){
  income_edu_2018[i,1] = sum(edu_income_range_2018[i,1:4]) # 0 - 49999 Low Level
  income_edu_2018[i,2] = sum(edu_income_range_2018[i,5:8]) # 49999- 99999 Medium Level
  income_edu_2018[i,3] = sum(edu_income_range_2018[i,9]) # >= 199999 High Level
}
income_edu_2018= as.data.frame(income_edu_2018)
colnames(income_edu_2018) = c('Low', 'Medium', 'High')
rownames(income_edu_2018) =c('9th_Grade', '12th_Grade', 'High_School', 'College', 'Associate','Bachelor')
income_edu_2018_mod = matrix(0, 3, 3)
for (i in 1:dim(income_edu_2018_mod)[2]){
  income_edu_2018_mod[1,i] = sum(income_edu_2018[1:3,i])
  income_edu_2018_mod[2,i] = sum(income_edu_2018[4:5,i])
  income_edu_2018_mod[3,i] = sum(income_edu_2018[6:9,i])
}
income_edu_2018_mod = as.data.frame(income_edu_2018_mod)
colnames(income_edu_2018_mod) = c('Low', 'Medium', "High")
rownames(income_edu_2018_mod) = c('HighSchool', 'College', 'Bachelor Above')

# Education 2019
Education_2019 <- read.csv("~/Desktop/VTCourse/STAT 5014/Group Project/Total work experience-both sexes.csv")
Education_2019 <- Education_2019[-c(1,2,8),]
Education_2019[,1] <- c('9th_Grade', '12th_Grade', 'High_School', 'College', 'Associate','Bachelor', 'Master')
rowname_edu = Education_2019[,1]
Education_2019 <- as.data.frame(Education_2019[,-(1:3)])
rownames(Education_2019) <- rowname_edu
colnames(Education_2019) = c('$2499', '$2500-$4999', '$5000-$7499', '$7500-$9999',
                             '$10000-$12499', '$12500-$14999', '$15000-$17499', '$17500-$19999', '$20000-$22499',
                             '$22500-$24999', '$25000-$27499', '$27500-$29999', '$30000-$32499', '$32500-$37499',
                             '$42500-$44999', '$45000-$47499', '$47999-$49999', '$50000-$52499', '$52500-$54999',
                             '$100000-')

edu_income_range_2019 = matrix(0, 9, 9)
for (i in 1:dim(Education_2019)[1]){
  for (j in 1:dim(Education_2019)[2]){
    Education_2019[i,j] = as.numeric(gsub("-", "", Education_2019[i,j]))
  }
}
for (i in 1:dim(edu_income_range_2019)[1]){
  edu_income_range_2019[i,1] = sum(as.numeric(Education_2019[i,1:5])) # 0 - 12499
  edu_income_range_2019[i,2] = sum(as.numeric(Education_2019[i,6:10])) #12500 - 24999
  edu_income_range_2019[i,3] = sum(as.numeric(Education_2019[i,11:15])) # 25000 - 37499
  edu_income_range_2019[i,4] = sum(as.numeric(Education_2019[i,16:20])) # 37500 - 49999
  edu_income_range_2019[i,5] = sum(as.numeric(Education_2019[i,21:25])) # 50000 - 62499
  edu_income_range_2019[i,6] = sum(as.numeric(Education_2019[i,26:30])) # 62500 - 74999
  edu_income_range_2019[i,7] = sum(as.numeric(Education_2019[i,31:35])) # 75000- 87499
  edu_income_range_2019[i,8] = sum(as.numeric(Education_2019[i,36:40])) # 87500 - 99999
  edu_income_range_2019[i,9] = sum(as.numeric(Education_2019[i,41])) # >= 100000
}
edu_income_range_2019 = as.data.frame(edu_income_range_2019)
rownames(edu_income_range_2019) = c('9th_Grade', '12th_Grade', 'High_School', 'College', 'Associate', 'Bachelor')
#colnames(range_combine_allr) = c('0to12499', '12500to24999', '25000to37499', '37500to49999', '50000to62499', '62500to74999', '75000to87499', '87500to99999', '100000to1000000')
colnames(edu_income_range_2019) = seq(0,100000, 12500)
# After divide the total income range data, I tried to divide it into low, medium and high level

```

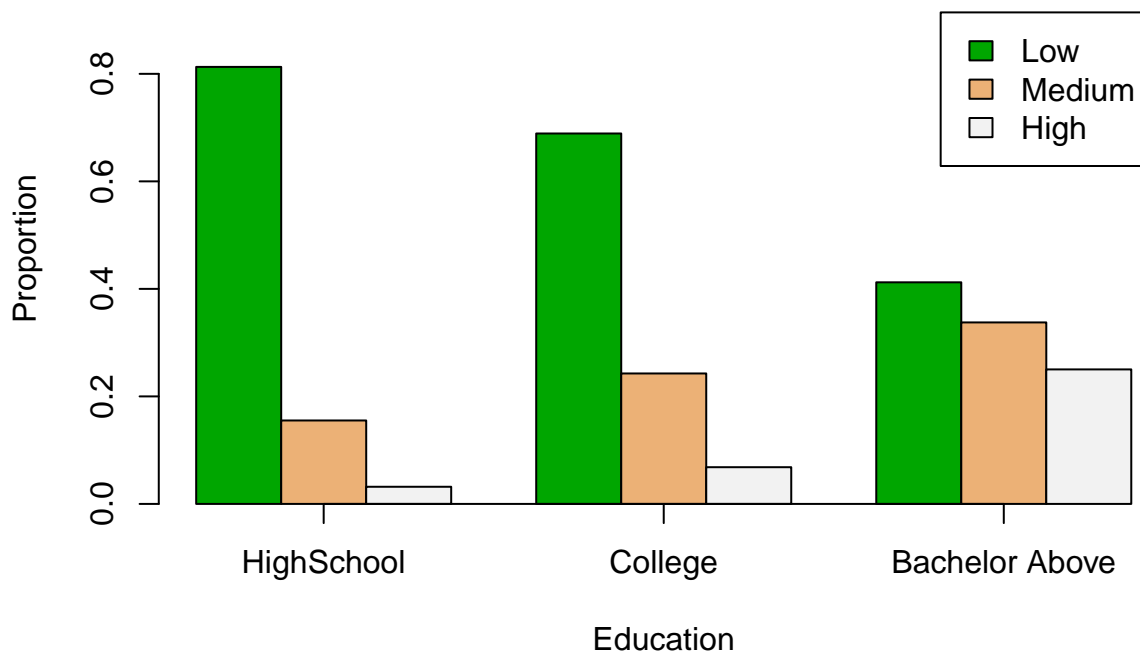
```

income_edu_2019 = matrix(0, 9, 3)
for (i in 1:9){
  income_edu_2019[i,1] = sum(edu_income_range_2019[i,1:4]) # 0 - 49999 Low Level
  income_edu_2019[i,2] = sum(edu_income_range_2019[i,5:8]) # 49999- 99999 Medium Level
  income_edu_2019[i,3] = sum(edu_income_range_2019[i,9]) # >= 199999 High Level
}
income_edu_2019 = as.data.frame(income_edu_2019)
colnames(income_edu_2019) = c('Low', 'Medium', 'High')
rownames(income_edu_2019) = c('9th_Grade', '12th_Grade', 'High_School', 'College', 'Associate', 'Bachelor')
income_edu_2019_mod = matrix(0, 3, 3)
for (i in 1:dim(income_edu_2019_mod)[2]){
  income_edu_2019_mod[1,i] = sum(income_edu_2019[1:3,i])
  income_edu_2019_mod[2,i] = sum(income_edu_2019[4:5,i])
  income_edu_2019_mod[3,i] = sum(income_edu_2019[6:9,i])}
income_edu_2019_mod = as.data.frame(income_edu_2019_mod)
colnames(income_edu_2019_mod) = c('Low', 'Medium', "High")
rownames(income_edu_2019_mod) = c('HighSchool', 'College', 'Bachelor Above')

income_edu_2017_mod = as.matrix(income_edu_2017_mod)
prop_income_edu_2017 = matrix(0, 3, 3)
for (i in 1:dim(income_edu_2017_mod)[1]){
  prop_income_edu_2017[i,1] = income_edu_2017_mod[i,1]/sum(income_edu_2017_mod[i,])
  prop_income_edu_2017[i,2] = income_edu_2017_mod[i,2]/sum(income_edu_2017_mod[i,])
  prop_income_edu_2017[i,3] = income_edu_2017_mod[i,3]/sum(income_edu_2017_mod[i,])
}
labels = rownames(income_edu_2017_mod)
color.names = terrain.colors(3)
barplot(t(prop_income_edu_2017),beside=T,ylim=c(0,0.95), col= color.names,xlab='Education',ylab="Proportion")

```

Income Level Versus Education of Year 2017

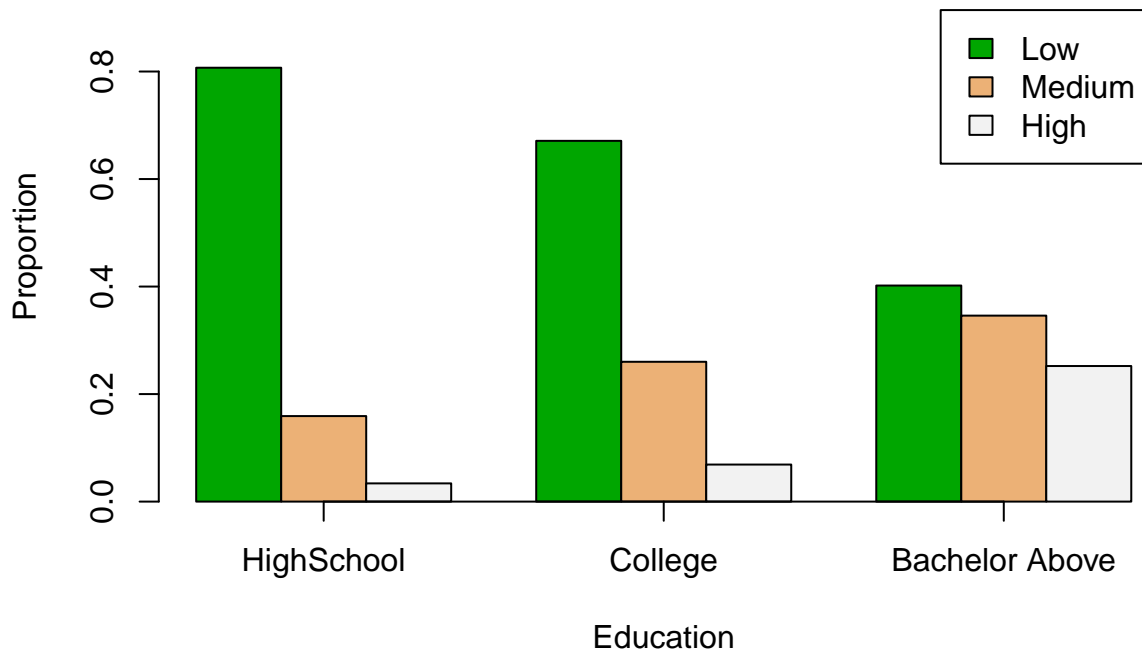


```

income_edu_2018_mod = as.matrix(income_edu_2018_mod)
prop_income_edu_2018 = matrix(0, 3, 3)
for (i in 1:dim(income_edu_2018_mod)[1]){
  prop_income_edu_2018[i,1] = income_edu_2018_mod[i,1]/sum(income_edu_2018_mod[i,])
  prop_income_edu_2018[i,2] = income_edu_2018_mod[i,2]/sum(income_edu_2018_mod[i,])
  prop_income_edu_2018[i,3] = income_edu_2018_mod[i,3]/sum(income_edu_2018_mod[i,])
}
labels = rownames(income_edu_2018_mod)
color.names = terrain.colors(3)
barplot(t(prop_income_edu_2018),beside=T,ylim=c(0,0.95), col= color.names,xlab='Education',ylab="Proportion")

```

Income Level Versus Education of Year 2018

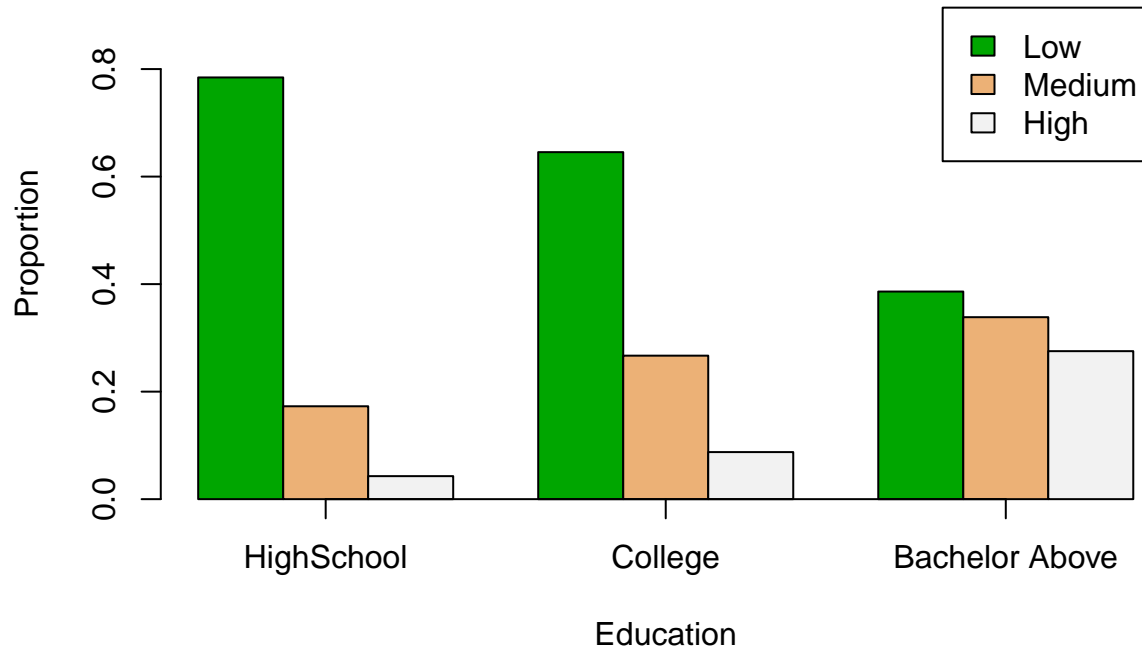


```

income_edu_2019_mod = as.matrix(income_edu_2019_mod)
prop_income_edu_2019 = matrix(0, 3, 3)
for (i in 1:dim(income_edu_2019_mod)[1]){
  prop_income_edu_2019[i,1] = income_edu_2019_mod[i,1]/sum(income_edu_2019_mod[i,])
  prop_income_edu_2019[i,2] = income_edu_2019_mod[i,2]/sum(income_edu_2019_mod[i,])
  prop_income_edu_2019[i,3] = income_edu_2019_mod[i,3]/sum(income_edu_2019_mod[i,])
}
labels = rownames(income_edu_2019_mod)
color.names = terrain.colors(3)
barplot(t(prop_income_edu_2019),beside=T,ylim=c(0,0.95), col= color.names,xlab='Education',ylab="Proportion")

```

Income Level Versus Education of Year 2019



```
Age_2019 <- read.csv("~/Desktop/VTCourse/STAT 5014/Group Project/Total work experience-both sexes/2019/
ind = c(1, 2, 3, 6,7, 9, 10, 12, 13, 15, 16,18, 19, 20 ,21)
Age_2019 = Age_2019[-ind,]
Age_2019_mod = as.data.frame(Age_2019[, -c(1,2,3)])
rownames(Age_2019_mod) = c('15 to 24', '25 to 35', '35 to 44', '45 to 54', '55 to 64', 'above 65')
colnames(Age_2019_mod) = c('$2499', '$2500-$4999', '$5000-$7499', '$7500-$9999',
                           '$10000-$12499', '$12500-$14999', '$15000-$17499', '$17500-$19999', '$20000-$22499',
                           '$22500-$24999', '$25000-$27499', '$27500-$29999', '$30000-$32499', '$32500-$37499',
                           '$42500-$44999', '$45000-$47499', '$47999-$49999', '$50000-$52499', '$52500-$54999',
                           '$100000-')
age_2019 = matrix(0, 6, 9)
for (i in 1:dim(Age_2019_mod)[1]){
  for (j in 1:dim(Age_2019_mod)[2]){
    Age_2019_mod[i,j] = as.numeric(gsub("-", "", Age_2019_mod[i,j]))
  }
}
for (i in 1:dim(age_2019)[1]){
  age_2019[i,1] = sum(as.numeric(Age_2019_mod[i,1:5])) # 0 - 12499
  age_2019[i,2] = sum(as.numeric(Age_2019_mod[i,6:10])) #12500 - 24999
  age_2019[i,3] = sum(as.numeric(Age_2019_mod[i,11:15])) # 25000 - 37499
  age_2019[i,4] = sum(as.numeric(Age_2019_mod[i,16:20])) # 37500 - 49999
  age_2019[i,5] = sum(as.numeric(Age_2019_mod[i,21:25])) # 50000 - 62499
  age_2019[i,6] = sum(as.numeric(Age_2019_mod[i,26:30])) # 62500 - 74999
  age_2019[i,7] = sum(as.numeric(Age_2019_mod[i,31:35])) # 75000- 87499
  age_2019[i,8] = sum(as.numeric(Age_2019_mod[i,36:40])) # 87500 - 99999
  age_2019[i,9] = sum(as.numeric(Age_2019_mod[i,41])) # >= 100000
}
age_2019 = as.data.frame(age_2019)
rownames(age_2019) = c('15 to 24', '25 to 35', '35 to 44', '45 to 54', '55 to 64', 'above 65')
Age_2019 = matrix(0, 6, 3)
```

```

for (i in 1:dim(Age_2019)[1]){
Age_2019[i,1] = sum(age_2019[i,1:3])
Age_2019[i,2] = sum(age_2019[i,4:6])
Age_2019[i,3] = sum(age_2019[i,6:9])
}
Age_2019 = as.data.frame(Age_2019)
colnames(Age_2019) = c('Low', 'Medium', "High")
rownames(Age_2019) =c('15 to 24', '25 to 35', '35 to 44', '45 to 54', '55 to 64', 'above 65')

# Age 2018
Age_2018 <- read.csv("~/Desktop/VTCourse/STAT 5014/Group Project/Total work experience-both sexes/2018/")
ind = c(1, 2, 3, 6,7, 9, 10, 12, 13, 15, 16,18, 19, 20 ,21)
Age_2018 = Age_2018[-ind,]
Age_2018_mod = as.data.frame(Age_2018[, -c(1,2,3)])
rownames(Age_2018_mod) = c('15 to 24', '25 to 35', '35 to 44', '45 to 54', '55 to 64', 'above 65')
colnames(Age_2018_mod) = c('$2499', '$2500-$4999', '$5000-$7499', '$7500-$9999',
'$10000-$12499', '$12500-$14999', '$15000-$17499', '$17500-$19999', '$20000-$22499',
'$22500-$24999', '$25000-$27499', '$27500-$29999', '$30000-$32499', '$32500-$37499',
'$42500-$44999', '$45000-$47499', '$47999-$49999', '$50000-$52499', '$52500-$54999',
'$100000-')
age_2018 = matrix(0, 6, 9)
for (i in 1:dim(Age_2018_mod)[1]){
  for (j in 1:dim(Age_2018_mod)[2]){
Age_2018_mod[i,j] = as.numeric(gsub("-", "", Age_2018_mod[i,j]))
  }
}
for (i in 1:dim(age_2018)[1]){
  age_2018[i,1] = sum(as.numeric(Age_2018_mod[i,1:5])) # 0 - 12499
  age_2018[i,2] = sum(as.numeric(Age_2018_mod[i,6:10])) #12500 - 24999
  age_2018[i,3] = sum(as.numeric(Age_2018_mod[i,11:15])) # 25000 - 37499
  age_2018[i,4] = sum(as.numeric(Age_2018_mod[i,16:20])) # 37500 - 49999
  age_2018[i,5] = sum(as.numeric(Age_2018_mod[i,21:25])) # 50000 - 62499
  age_2018[i,6] = sum(as.numeric(Age_2018_mod[i,26:30])) # 62500 - 74999
  age_2018[i,7] = sum(as.numeric(Age_2018_mod[i,31:35])) # 75000- 87499
  age_2018[i,8] = sum(as.numeric(Age_2018_mod[i,36:40])) # 87500 - 99999
  age_2018[i,9] = sum(as.numeric(Age_2018_mod[i,41])) # >= 100000
}
age_2018 = as.data.frame(age_2018)
rownames(age_2018) =c('15 to 24', '25 to 35', '35 to 44', '45 to 54', '55 to 64', 'above 65')
Age_2018 = matrix(0, 6, 3)
for (i in 1:dim(Age_2018)[1]){
Age_2018[i,1] = sum(age_2018[i,1:3])
Age_2018[i,2] = sum(age_2018[i,4:6])
Age_2018[i,3] = sum(age_2018[i,6:9])
}
Age_2018 = as.data.frame(Age_2018)
colnames(Age_2018) = c('Low', 'Medium', "High")
rownames(Age_2018) =c('15 to 24', '25 to 35', '35 to 44', '45 to 54', '55 to 64', 'above 65')

# Age 2017
Age_2017 <- read.csv("~/Desktop/VTCourse/STAT 5014/Group Project/Total work experience-both sexes/2017/")
ind = c(1, 2, 3, 6,7, 9, 10, 12, 13, 15, 16,18, 19, 20 ,21)
Age_2017 = Age_2017[-ind,]
Age_2017_mod = as.data.frame(Age_2017[, -c(1,2,3)])

```

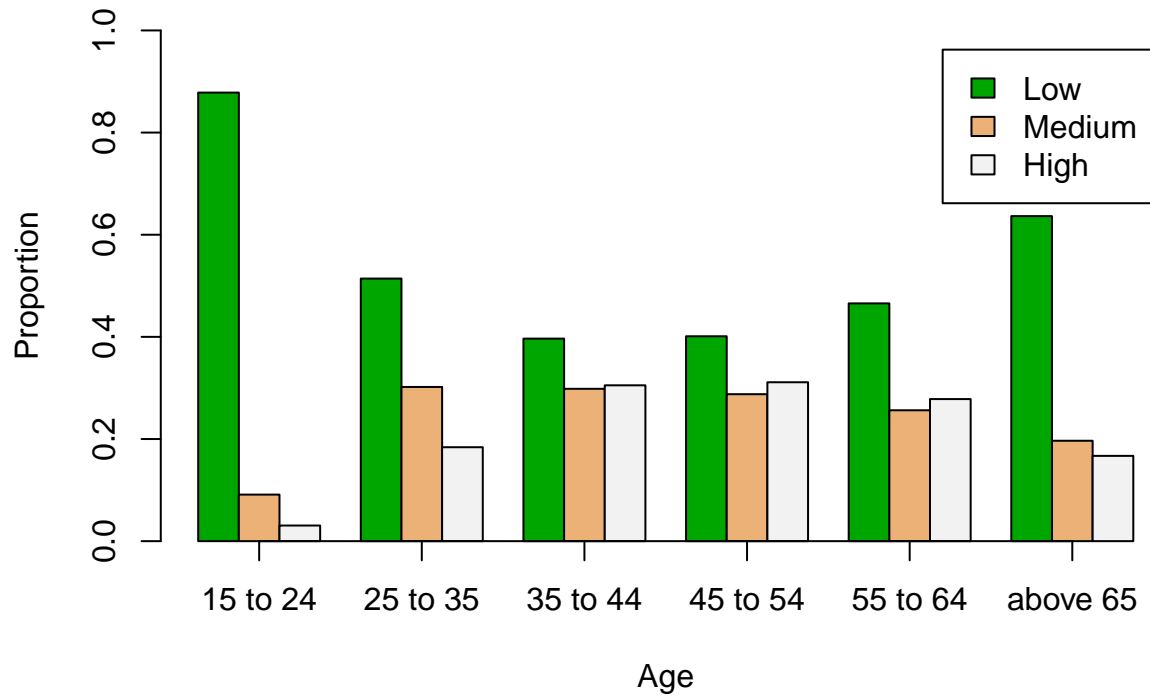
```

rownames(Age_2017_mod) = c('15 to 24', '25 to 35', '35 to 44', '45 to 54', '55 to 64', 'above 65')
colnames(Age_2017_mod) = c('$2499', '$2500-$4999', '$5000-$7499', '$7500-$9999',
                            '$10000-$12499', '$12500-$14999', '$15000-$17499', '$17500-$19999', '$20000-$22499',
                            '$22500-$24999', '$25000-$27499', '$27500-$29999', '$30000-$32499', '$32500-$37499',
                            '$42500-$44999', '$45000-$47499', '$47999-$49999', '$50000-$52499', '$52500-$54999',
                            '$100000-')
age_2017 = matrix(0, 6, 9)
for (i in 1:dim(Age_2017_mod)[1]){
  for (j in 1:dim(Age_2017_mod)[2]){
    Age_2017_mod[i,j] = as.numeric(gsub("-", "", Age_2017_mod[i,j]))
  }
}
for (i in 1:dim(age_2017)[1]){
  age_2017[i,1] = sum(as.numeric(Age_2017_mod[i,1:5])) # 0 - 12499
  age_2017[i,2] = sum(as.numeric(Age_2017_mod[i,6:10])) # 12500 - 24999
  age_2017[i,3] = sum(as.numeric(Age_2017_mod[i,11:15])) # 25000 - 37499
  age_2017[i,4] = sum(as.numeric(Age_2017_mod[i,16:20])) # 37500 - 49999
  age_2017[i,5] = sum(as.numeric(Age_2017_mod[i,21:25])) # 50000 - 62499
  age_2017[i,6] = sum(as.numeric(Age_2017_mod[i,26:30])) # 62500 - 74999
  age_2017[i,7] = sum(as.numeric(Age_2017_mod[i,31:35])) # 75000 - 87499
  age_2017[i,8] = sum(as.numeric(Age_2017_mod[i,36:40])) # 87500 - 99999
  age_2017[i,9] = sum(as.numeric(Age_2017_mod[i,41])) # >= 100000
}
age_2017 = as.data.frame(age_2017)
rownames(age_2017) = c('15 to 24', '25 to 35', '35 to 44', '45 to 54', '55 to 64', 'above 65')
Age_2017 = matrix(0, 6, 3)
for (i in 1:dim(Age_2017)[1]){
  Age_2017[i,1] = sum(age_2017[i,1:3])
  Age_2017[i,2] = sum(age_2017[i,4:6])
  Age_2017[i,3] = sum(age_2017[i,6:9])
}
Age_2017 = as.data.frame(Age_2017)
colnames(Age_2017) = c('Low', 'Medium', 'High')
rownames(Age_2017) = c('15 to 24', '25 to 35', '35 to 44', '45 to 54', '55 to 64', 'above 65')

Age_2017 = as.matrix(Age_2017)
prop_age_2017 = matrix(0, 6, 3)
for (i in 1:dim(prop_age_2017)[1]){
  prop_age_2017[i,1] = Age_2017[i,1]/sum(Age_2017[i,])
  prop_age_2017[i,2] = Age_2017[i,2]/sum(Age_2017[i,])
  prop_age_2017[i,3] = Age_2017[i,3]/sum(Age_2017[i,])
}
labels = rownames(Age_2017)
color.names = terrain.colors(3)
barplot(t(prop_age_2017), beside=T, ylim=c(0,1), col= color.names,
        xlab='Age', ylab="Proportion", axis.lty="solid", legend = colnames(Age_2017), names.arg=labels, m

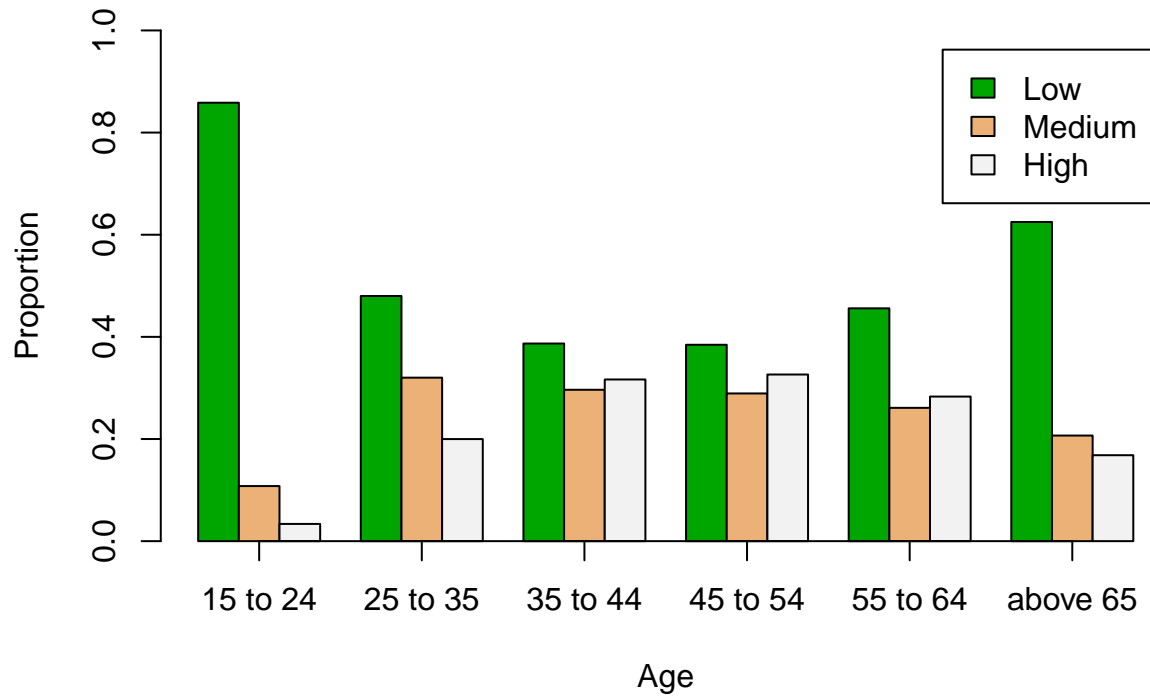
```


Income Level Versus Age of Year 2017



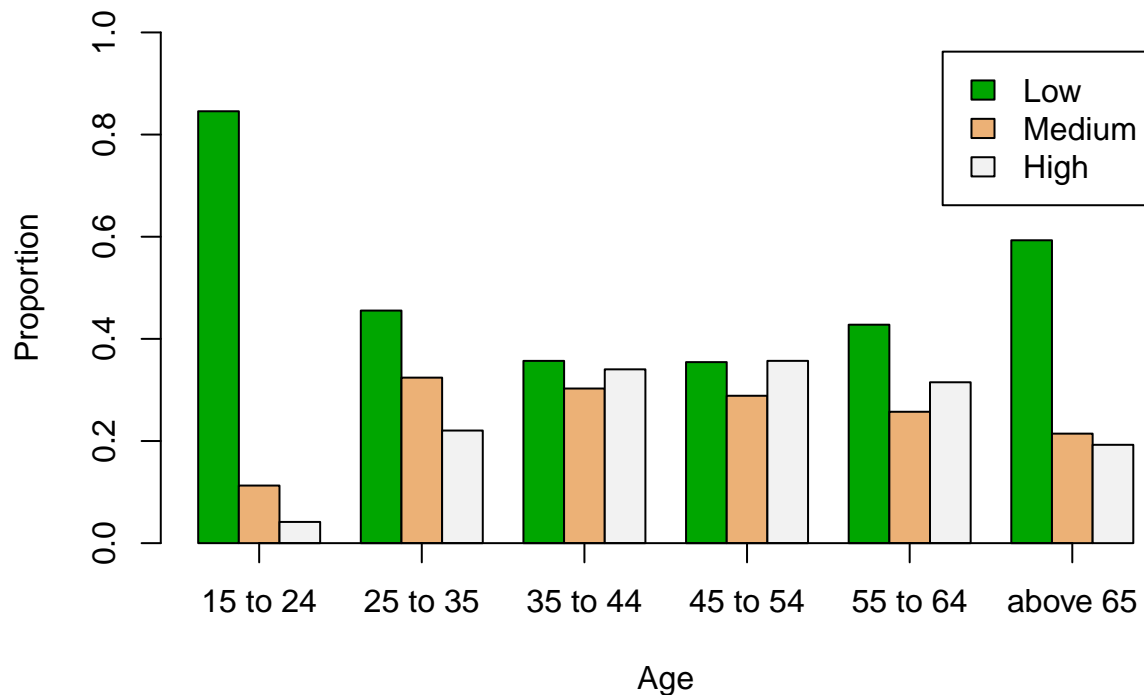
```
Age_2018 = as.matrix(Age_2018)
prop_age_2018 = matrix(0, 6, 3)
for (i in 1:dim(prop_age_2018)[1]){
  prop_age_2018[i,1] = Age_2018[i,1]/sum(Age_2018[i,])
  prop_age_2018[i,2] = Age_2018[i,2]/sum(Age_2018[i,])
  prop_age_2018[i,3] = Age_2018[i,3]/sum(Age_2018[i,])
}
labels = rownames(Age_2018)
color.names = terrain.colors(3)
barplot(t(prop_age_2018), beside=T,ylim=c(0,1), col= color.names,
        xlab='Age',ylab="Proportion",axis.lty="solid", legend = colnames(Age_2018), names.arg=labels, m
```

Income Level Versus Age of Year 2018



```
Age_2019 = as.matrix(Age_2019)
prop_age_2019 = matrix(0, 6, 3)
for (i in 1:dim(prop_age_2019)[1]){
  prop_age_2019[i,1] = Age_2019[i,1]/sum(Age_2019[i,])
  prop_age_2019[i,2] = Age_2019[i,2]/sum(Age_2019[i,])
  prop_age_2019[i,3] = Age_2019[i,3]/sum(Age_2019[i,])
}
labels = rownames(Age_2019)
color.names = terrain.colors(3)
barplot(t(prop_age_2019), beside=T,ylim=c(0,1), col= color.names,
        xlab='Age',ylab="Proportion",axis.lty="solid", legend = colnames(Age_2019), names.arg=labels, m
```

Income Level Versus Age of Year 2019



2016 region

```
allraces_2016 <- read.csv("~/Desktop/VTCourse/STAT 5014/Group Project/Total work experience-both sexes/")
allraces_2016 = as.data.frame(allraces_2016[-1,])
colnames(allraces_2016) = c('Region', 'Total', 'Total with Income', '$2499', '$2500-$4999', '$5000-$7499',
'$10000-$12499', '$12500-$14999', '$15000-$17499', '$17500-$19999', '$20000-$22499',
'$22500-$24999', '$25000-$27499', '$27500-$29999', '$30000-$32499', '$32500-$37499',
'$42500-$44999', '$45000-$47499', '$47500-$49999', '$50000-$52499', '$52500-$54999',
'$100000-')
)
all_race_income_range_2016 = allraces_2016[,4:44]
colnames(all_race_income_range_2016) = c('$2499', '$2500-$4999', '$5000-$7499', '$7500-$9999',
'$10000-$12499', '$12500-$14999', '$15000-$17499', '$17500-$19999', '$20000-$22499',
'$22500-$24999', '$25000-$27499', '$27500-$29999', '$30000-$32499', '$32500-$37499',
'$42500-$44999', '$45000-$47499', '$47999-$49999', '$50000-$52499', '$52500-$54999',
'$100000-')
range_combine_allr_2016 = matrix(0, 4, 9)
for (i in 1:dim(all_race_income_range_2016)[1]){
  for (j in 1:dim(all_race_income_range_2016)[2]){
    all_race_income_range_2016[i,j] = as.numeric(gsub("-", "", all_race_income_range_2016[i,j]))
  }
}
for (i in 1:dim(all_race_income_range_2016)[1]){
  range_combine_allr_2016[i,1] = sum(as.numeric(all_race_income_range_2016[i,1:5])) # 0 - 12499
  range_combine_allr_2016[i,2] = sum(as.numeric(all_race_income_range_2016[i,6:10])) # 12500 - 24999
  range_combine_allr_2016[i,3] = sum(as.numeric(all_race_income_range_2016[i,11:15])) # 25000 - 37499
  range_combine_allr_2016[i,4] = sum(as.numeric(all_race_income_range_2016[i,16:20])) # 37500 - 49999
  range_combine_allr_2016[i,5] = sum(as.numeric(all_race_income_range_2016[i,21:25])) # 50000 - 62499
  range_combine_allr_2016[i,6] = sum(as.numeric(all_race_income_range_2016[i,26:30])) # 62500 - 74999
  range_combine_allr_2016[i,7] = sum(as.numeric(all_race_income_range_2016[i,31:35])) # 75000 - 87499
```

```

    range_combine_allr_2016[i,8] = sum(as.numeric(all_race_income_range_2016[i,36:40])) # 87500 - 99999
    range_combine_allr_2016[i,9] = sum(as.numeric(all_race_income_range_2016[i,41])) # >= 100000
  }
range_combine_allr_2016 = as.data.frame(range_combine_allr_2016)
rownames(range_combine_allr_2016) = c('Northeast', 'Midwest', 'South', 'West')
#colnames(range_combine_allr) = c('0to12499', '12500to24999', '25000to37499', '37500to49999', '50000to62499', '62500to74999', '75000to87499', '87500to99999', '100000to')
colnames(range_combine_allr_2016) = seq(0,100000, 12500)
# After divide the total income range data, I tried to divide it into low, medium and high level
income_level_all_2016 = matrix(0, 4, 3)
for (i in 1:4){
  income_level_all_2016[i,1] = sum(range_combine_allr_2016[i,1:4]) # 0 - 49999 Low Level
  income_level_all_2016[i,2] = sum(range_combine_allr_2016[i,5:8]) # 49999- 99999 Medium Level
  income_level_all_2016[i,3] = sum(range_combine_allr_2016[i,9]) # >= 199999 High Level
}
income_level_all_2016= as.data.frame(income_level_all_2016)
colnames(income_level_all_2016) = c('Low', 'Medium', 'High')
rownames(income_level_all_2016) = c('Northeast', 'Midwest', 'South', 'West')

# 2015 region
allraces_2015 <- read.csv("~/Desktop/VTCourse/STAT 5014/Group Project/Total work experience-both sexes/")
allraces_2015 = as.data.frame(allraces_2015[-1,])
colnames(allraces_2015) = c('Region','Total', 'Total with Income', '$2499', '$2500-$4999', '$5000-$7499', '$7500-$9999', '$10000-$12499', '$12500-$14999', '$15000-$17499', '$17500-$19999', '$20000-$22499', '$22500-$24999', '$25000-$27499', '$27500-$29999', '$30000-$32499', '$32500-$37499', '$37500-$49999', '$42500-$44999', '$45000-$47499', '$47500-$49999', '$50000-$52499', '$52500-$54999', '$55000-$57499', '$57500-$59999', '$60000-$62499', '$62500-$64999', '$65000-$67499', '$67500-$69999', '$70000-$72499', '$72500-$74999', '$75000-$77499', '$77500-$79999', '$80000-$82499', '$82500-$84999', '$85000-$87499', '$87500-$89999', '$90000-$92499', '$92500-$94999', '$95000-$97499', '$97500-$99999', '$100000-')
)
all_race_income_range_2015 = allraces_2015[,4:44]
colnames(all_race_income_range_2015) = c('$2499', '$2500-$4999', '$5000-$7499', '$7500-$9999', '$10000-$12499', '$12500-$14999', '$15000-$17499', '$17500-$19999', '$20000-$22499', '$22500-$24999', '$25000-$27499', '$27500-$29999', '$30000-$32499', '$32500-$37499', '$37500-$49999', '$42500-$44999', '$45000-$47499', '$47500-$49999', '$50000-$52499', '$52500-$54999', '$55000-$57499', '$57500-$59999', '$60000-$62499', '$62500-$64999', '$65000-$67499', '$67500-$69999', '$70000-$72499', '$72500-$74999', '$75000-$77499', '$77500-$79999', '$80000-$82499', '$82500-$84999', '$85000-$87499', '$87500-$89999', '$90000-$92499', '$92500-$94999', '$95000-$97499', '$97500-$99999', '$100000-')
range_combine_allr_2015 = matrix(0, 4, 9)
for (i in 1:dim(all_race_income_range_2015)[1]){
  for (j in 1:dim(all_race_income_range_2015)[2]){
    all_race_income_range_2015[i,j] = as.numeric(gsub("-", "", all_race_income_range_2015[i,j]))
  }
}
for (i in 1:dim(all_race_income_range_2015)[1]){
  range_combine_allr_2015[i,1] = sum(as.numeric(all_race_income_range_2015[i,1:5])) # 0 - 12499
  range_combine_allr_2015[i,2] = sum(as.numeric(all_race_income_range_2015[i,6:10])) #12500 - 24999
  range_combine_allr_2015[i,3] = sum(as.numeric(all_race_income_range_2015[i,11:15])) # 25000 - 37499
  range_combine_allr_2015[i,4] = sum(as.numeric(all_race_income_range_2015[i,16:20])) # 37500 - 49999
  range_combine_allr_2015[i,5] = sum(as.numeric(all_race_income_range_2015[i,21:25])) # 50000 - 62499
  range_combine_allr_2015[i,6] = sum(as.numeric(all_race_income_range_2015[i,26:30])) # 62500 - 74999
  range_combine_allr_2015[i,7] = sum(as.numeric(all_race_income_range_2015[i,31:35])) # 75000- 87499
  range_combine_allr_2015[i,8] = sum(as.numeric(all_race_income_range_2015[i,36:40])) # 87500 - 99999
  range_combine_allr_2015[i,9] = sum(as.numeric(all_race_income_range_2015[i,41])) # >= 100000
}
range_combine_allr_2015 = as.data.frame(range_combine_allr_2015)
rownames(range_combine_allr_2015) = c('Northeast', 'Midwest', 'South', 'West')
#colnames(range_combine_allr) = c('0to12499', '12500to24999', '25000to37499', '37500to49999', '50000to62499', '62500to74999', '75000to87499', '87500to99999', '100000to')
colnames(range_combine_allr_2015) = seq(0,100000, 12500)

```

```

# After divide the total income range data, I tried to divide it into low, medium and high level
income_level_all_2015 = matrix(0, 4, 3)
for (i in 1:4){
  income_level_all_2015[i,1] = sum(range_combine_allr_2015[i,1:4]) # 0 - 49999 Low Level
  income_level_all_2015[i,2] = sum(range_combine_allr_2015[i,5:8]) # 49999- 99999 Medium Level
  income_level_all_2015[i,3] = sum(range_combine_allr_2015[i,9]) # >= 199999 High Level
}
income_level_all_2015= as.data.frame(income_level_all_2015)
colnames(income_level_all_2015) = c('Low', 'Medium', 'High')
rownames(income_level_all_2015) = c('Northeast', 'Midwest', 'South', 'West')

```

For one region, time series data analysis:

```

# Northeast Region Past five years
Northeast = rbind(income_level_all_2015[1,],income_level_all_2016[1,],income_level_all_2017[1,],income_level_all_2018[1,],income_level_all_2019[1,])
rownames(Northeast) = c('2015', '2016', '2017', '2018', '2019')
# Midwest Region Past five years
Midwest = rbind(income_level_all_2015[2,],income_level_all_2016[2,],income_level_all_2017[2,],income_level_all_2018[2,],income_level_all_2019[2,])
rownames(Midwest) = c('2015', '2016', '2017', '2018', '2019')
# South Region Past five years
South = rbind(income_level_all_2015[3,],income_level_all_2016[3,],income_level_all_2017[3,],income_level_all_2018[3,],income_level_all_2019[3,])
rownames(South) = c('2015', '2016', '2017', '2018', '2019')
# West Region Past Five years
West = rbind(income_level_all_2015[4,],income_level_all_2016[4,],income_level_all_2017[4,],income_level_all_2018[4,],income_level_all_2019[4,])
rownames(West) = c('2015', '2016', '2017', '2018', '2019')
prop <- function(data){
  dat_prop = matrix(0, 5, 3)
  for (i in 1:dim(data)[1]){
    dat_prop[i,1] = data[i,1]/sum(data[i,])
    dat_prop[i,2] = data[i,2]/sum(data[i,])
    dat_prop[i,3] = data[i,3]/sum(data[i,])
  }
  return(dat_prop)
}
# West
west_prop = as.data.frame(prop(West))
rownames(west_prop) = c('2015', '2016', '2017', '2018', '2019')
colnames(west_prop) = c('Low', 'Medium', 'High')
# Northeast
northeast_prop = as.data.frame(prop(Northeast))
rownames(northeast_prop) = c('2015', '2016', '2017', '2018', '2019')
colnames(northeast_prop) = c('Low', 'Medium', 'High')
# South
south_prop = as.data.frame(prop(South))
rownames(south_prop) = c('2015', '2016', '2017', '2018', '2019')
colnames(south_prop) = c('Low', 'Medium', 'High')
# Midwest
midwest_prop = as.data.frame(prop(Midwest))
rownames(midwest_prop) = c('2015', '2016', '2017', '2018', '2019')
colnames(midwest_prop) = c('Low', 'Medium', 'High')

par(mfrow=c(2,2))
# Midwest
Year = seq(2015,2019,1)

```

```

# plot the first curve by calling plot() function
# First curve is plotted
plot(Year, midwest_prop[,1], type="o", col="blue", pch="o", lty=1, ylim=c(0,0.7), main = 'Proportion of
# Add second curve to the same plot by calling points() and lines()
# Use symbol '*' for points.
points(Year,midwest_prop[,2], col="orange")
lines(Year,midwest_prop[,2], col="orange",lty=2)
# Add Third curve to the same plot by calling points() and lines()
# Use symbol '+' for points.
points(Year, midwest_prop[,3], col="black")
lines(Year, midwest_prop[,3], col="black", lty=3)
legend(2015, 0.6, legend=c("Low Level", "Medium Level",'High Level'),
      col=c("blue", "orange", 'black'),lty= 1:3, cex=0.6)

# Northeast
Year = seq(2015,2019,1)
# plot the first curve by calling plot() function
# First curve is plotted
plot(Year, northeast_prop[,1], type="o", col="purple", pch="o", lty=1, ylim=c(0,0.7), main = 'Proportion
# Add second curve to the same plot by calling points() and lines()
# Use symbol '*' for points.
points(Year,northeast_prop[,2], col="darkgreen")
lines(Year,northeast_prop[,2], col="darkgreen",lty=2)
# Add Third curve to the same plot by calling points() and lines()
# Use symbol '+' for points.
points(Year, northeast_prop[,3], col="black")
lines(Year, northeast_prop[,3], col='black', lty=3)
legend(2015, 0.6, legend=c("Low Level", "Medium Level",'High Level'),
      col=c("purple", "darkgreen", 'black'),lty= 1:3, cex=0.6)

# West
Year = seq(2015,2019,1)
# plot the first curve by calling plot() function
# First curve is plotted
plot(Year, west_prop[,1], type="o", col="green", pch="o", lty=1, ylim=c(0,0.7), main = 'Proportion of D
# Add second curve to the same plot by calling points() and lines()
# Use symbol '*' for points.
points(Year,west_prop[,2], col="gray")
lines(Year,west_prop[,2], col="gray",lty=2)
# Add Third curve to the same plot by calling points() and lines()
# Use symbol '+' for points.
points(Year, west_prop[,3], col="black")
lines(Year, west_prop[,3], col="black", lty=3)
legend(2015, 0.6, legend=c("Low Level", "Medium Level",'High Level'),
      col=c("green", "gray", 'black'),lty= 1:3, cex=0.6)

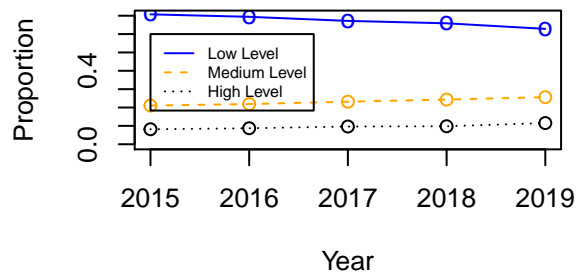
# South
Year = seq(2015,2019,1)
# plot the first curve by calling plot() function
# First curve is plotted
plot(Year, south_prop[,1], type="o", col="pink", pch="o", lty=1, ylim=c(0,0.7), main = 'Proportion of D
# Add second curve to the same plot by calling points() and lines()
# Use symbol '*' for points.
points(Year,south_prop[,2], col="brown")
lines(Year,south_prop[,2], col="brown",lty=2)

```

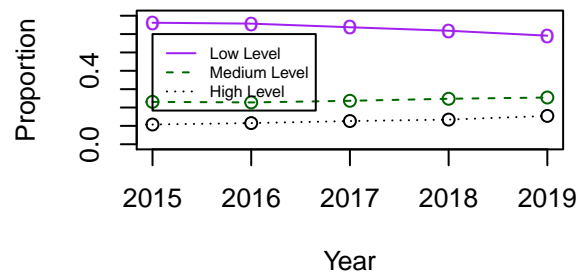


```
# Add Third curve to the same plot by calling points() and lines()
# Use symbol '+' for points.
points(Year, south_prop[,3], col="black")
lines(Year, south_prop[,3], col="black", lty=3)
legend(2015, 0.6, legend=c("Low Level", "Medium Level", "High Level"),
      col=c("pink", "brown", "black"), lty= 1:3, cex=0.6)
```

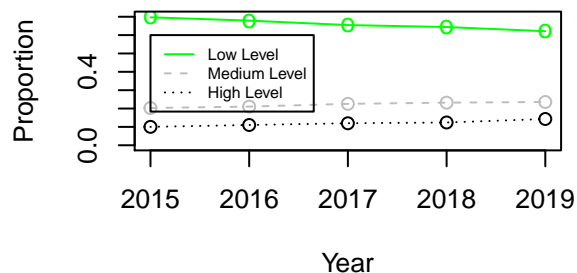
Proportion of Different Income Levels in Midwest Versus Year



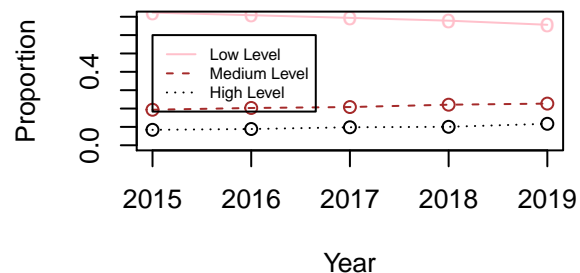
Proportion of Different Income Levels in Northeast Versus Year



Proportion of Different Income Levels in West Versus Year



Proportion of Different Income Levels in South Versus Year



```
# part regression versus time
# Northeast eg.
dat_noreast_low = as.data.frame(cbind(Year, northeast_prop[,1]))
colnames(dat_noreast_low) = c('Year', 'Prop_low')
fit_low = lm(Prop_low ~ Year, data = dat_noreast_low)
summary(fit_low)
```

```
##
## Call:
## lm(formula = Prop_low ~ Year, data = dat_noreast_low)
##
## Residuals:
##      1       2       3       4       5
## -0.007286  0.005938  0.004114  0.003102 -0.005868
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.861015   4.500148   8.191  0.00381 **
## Year        -0.017961   0.002231  -8.050  0.00400 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.007055 on 3 degrees of freedom
## Multiple R-squared: 0.9558, Adjusted R-squared: 0.941
## F-statistic: 64.81 on 1 and 3 DF, p-value: 0.004003

dat_noreast_medium = as.data.frame(cbind(Year, northeast_prop[,2]))
colnames(dat_noreast_medium) = c('Year', 'Prop_medium')
fit_medium = lm(Prop_medium ~ Year, data = dat_noreast_medium)
summary(fit_medium)

##
## Call:
## lm(formula = Prop_medium ~ Year, data = dat_noreast_medium)
##
## Residuals:
##      1      2      3      4      5
## 0.004764 -0.004946 -0.002829 0.001441 0.001571
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -13.330106   2.845577  -4.685   0.0184 *
## Year          0.006728   0.001411   4.769   0.0175 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.004461 on 3 degrees of freedom
## Multiple R-squared: 0.8834, Adjusted R-squared: 0.8446
## F-statistic: 22.74 on 1 and 3 DF, p-value: 0.01752

dat_noreast_high = as.data.frame(cbind(Year, northeast_prop[,3]))
colnames(dat_noreast_high) = c('Year', 'Prop_high')
fit_high = lm(Prop_high ~ Year, data = dat_noreast_high)
summary(fit_high)

##
## Call:
## lm(formula = Prop_high ~ Year, data = dat_noreast_high)
##
## Residuals:
##      1      2      3      4      5
## 0.0025221 -0.0009919 -0.0012851 -0.0045427 0.0042976
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -22.530909   2.554028  -8.822 0.00307 **
## Year          0.011234   0.001266   8.872 0.00302 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.004004 on 3 degrees of freedom
## Multiple R-squared: 0.9633, Adjusted R-squared: 0.951
## F-statistic: 78.71 on 1 and 3 DF, p-value: 0.00302

Low = as.data.frame(cbind(northeast_prop[,1], south_prop[,1], west_prop[,1], midwest_prop[,1]))
colnames(Low) = c('Northeast', 'South', 'West', 'Midwest')
rownames(Low) = c('2015', '2016', '2017', '2018', '2019')
t = as.data.frame(melt(Low))
```

```
## No id variables; using all as measure variables
fit_low = lm(value~variable, data=t)
summary(fit_low)

##
## Call:
## lm(formula = value ~ variable, data = t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.043739 -0.014959  0.000823  0.022421  0.037841
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.63311    0.01293  48.959 < 2e-16 ***
## variableSouth    0.05886    0.01829   3.218  0.00537 **
## variableWest     0.02603    0.01829   1.423  0.17390
## variableMidwest  0.03900    0.01829   2.133  0.04878 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02892 on 16 degrees of freedom
## Multiple R-squared:  0.4056, Adjusted R-squared:  0.2942
## F-statistic: 3.639 on 3 and 16 DF,  p-value: 0.03566

# Medium
Medium = as.data.frame( cbind(northeast_prop[,2], south_prop[,2], west_prop[,2], midwest_prop[,2]))
colnames(Medium) = c('Northeast', 'South', 'West', 'Midwest')
rownames(Medium) = c('2015', '2016', '2017', '2018', '2019')
m = as.data.frame(melt(Medium))

## No id variables; using all as measure variables
fit_medium = lm(value~variable, data=m)
summary(fit_medium)

##
## Call:
## lm(formula = value ~ variable, data = m)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.021369 -0.010960 -0.001403  0.010777  0.024083
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.239449    0.006494  36.873 < 2e-16 ***
## variableSouth  -0.028985    0.009184  -3.156  0.00612 **
## variableWest   -0.018078    0.009184  -1.968  0.06659 .
## variableMidwest -0.007283    0.009184  -0.793  0.43938
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01452 on 16 degrees of freedom
## Multiple R-squared:  0.4165, Adjusted R-squared:  0.3071
```

```
## F-statistic: 3.807 on 3 and 16 DF, p-value: 0.03107
# High
High = as.data.frame( cbind(northeast_prop[,3], south_prop[,3], west_prop[,3], midwest_prop[,3]))
colnames(High) = c('Northeast', 'South', 'West', 'Midwest')
rownames(High) = c('2015', '2016', '2017', '2018', '2019')
h = as.data.frame(melt(High))

## No id variables; using all as measure variables
fit_high = lm(value~variable, data=h)
summary(fit_high)

##
## Call:
## lm(formula = value ~ variable, data = h)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.0199453 -0.0098606  0.0003246  0.0051092  0.0267649
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.127442   0.006756  18.862 2.36e-12 ***
## variableSouth  -0.029870   0.009555  -3.126  0.00651 **
## variableWest   -0.007948   0.009555  -0.832  0.41773
## variableMidwest -0.031721   0.009555  -3.320  0.00433 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01511 on 16 degrees of freedom
## Multiple R-squared:  0.5075, Adjusted R-squared:  0.4152
## F-statistic: 5.496 on 3 and 16 DF, p-value: 0.008654
```