Home-Work(3)

```
Home-Work(3)


Section A
```

This section is for testing your data transformation, ggplot, and base function skills. If it's neccessary, please deal with the overplotting, or labels on axis/legend properly.

Problem 1

We would like to create one simulated (fake) data frame contained the employer height(cm) information from two companies: Alpha and Beta. Also, this data frame should include the companies' area codes. (Company may have multiple subsidiaries in different areas)

• Create one column Area Code with 2000 rows only contained 26 upper-case letters (alphabet). These letters should be randomly filled in 2000 rows. (with replacement)

• Create one column Company with 2000 rows contained only two values "Alpha" and "Beta". To be convenient, first 1000 rows should be "Alpha"s, and last 1000 rows should be "Beta"s.

• Create one column Employee Height (cm) with 2000 rows. To be convenient,first 1000 rows and last 1000 rows should be randomly generated with mean = 160, sd = 5, and mean = 170, sd = 5, respectively.

Then create a density plot on the height, mapping company as the fill. hints: The built-in "LETTERS" contains 26 upper-case letters.

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────────────────────────
──────── tidyverse 1.3.0 ──
```

```
## ✓ ggplot2 3.2.1     ✓ purrr   0.3.3
## ✓ tibble  2.1.3     ✓ dplyr   0.8.3
## ✓ tidyr   1.0.0     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.4.0
```

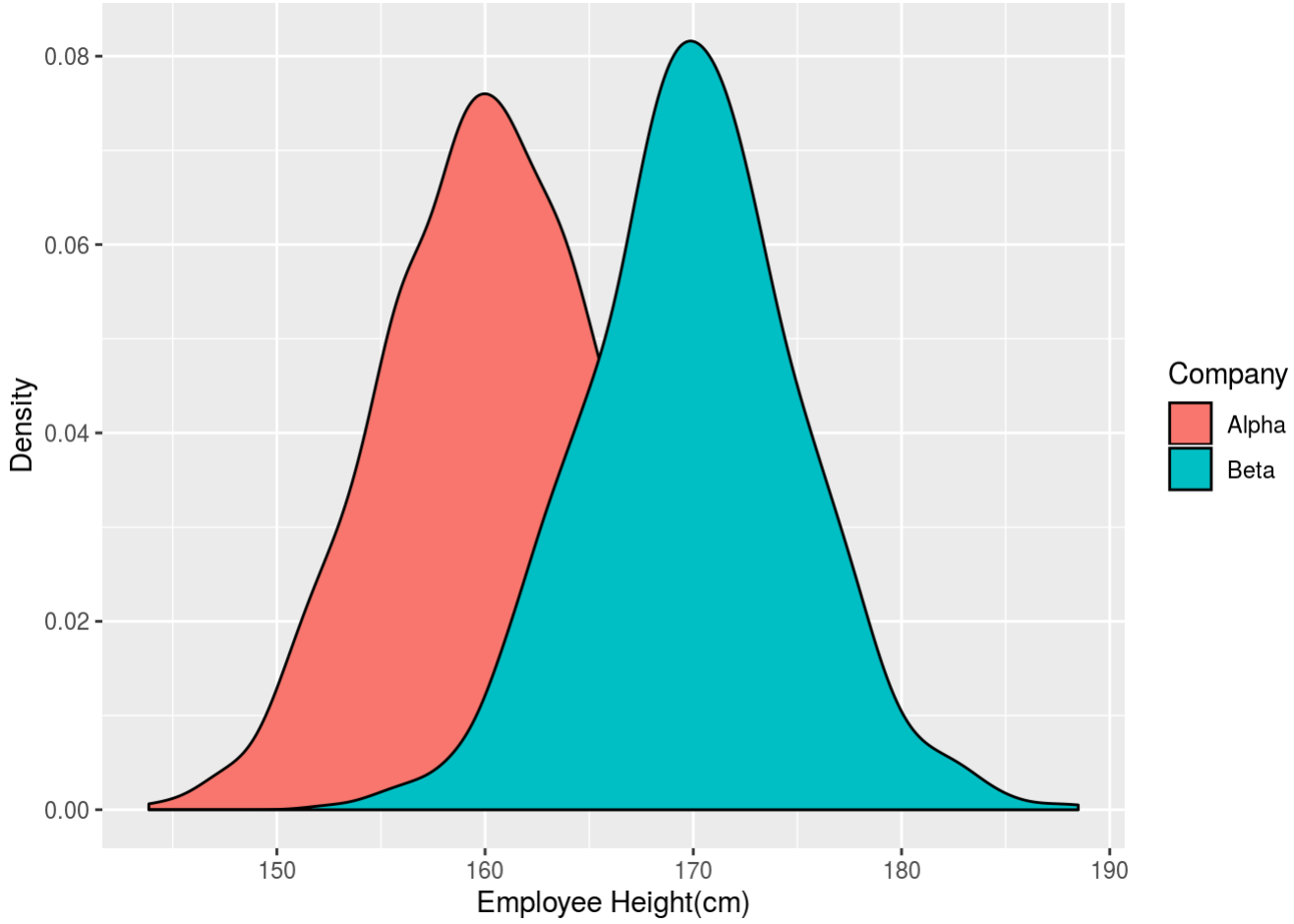
```
## ── Conflicts ────────────────────────────────────────────────────────────────
─── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(haven)
```

```
##### Creating a data frame
set.seed(10)
company_df<-data.frame(`Area code`=sample(LETTERS[1:26],2000,replace=TRUE),`Company`=rep(c('Alph
a','Beta'),each=1000),`Employee Height(cm)`=rnorm(2000, mean=rep(c(160,170),each=1000),sd=5),che
ck.names = F)
head(company_df)
```

```
##   Area code Company Employee Height(cm)
## 1         K   Alpha            162.9499
## 2         I   Alpha            157.7288
## 3         J   Alpha            158.9471
## 4         P   Alpha            164.6853
## 5         L   Alpha            157.0649
## 6         W   Alpha            162.4829
```

```
##### Creating a density plot
density_plot<-ggplot(company_df)+geom_density(mapping=aes(`Employee Height(cm)`,fill=`Company`))
+ylab('Density')
density_plot
```



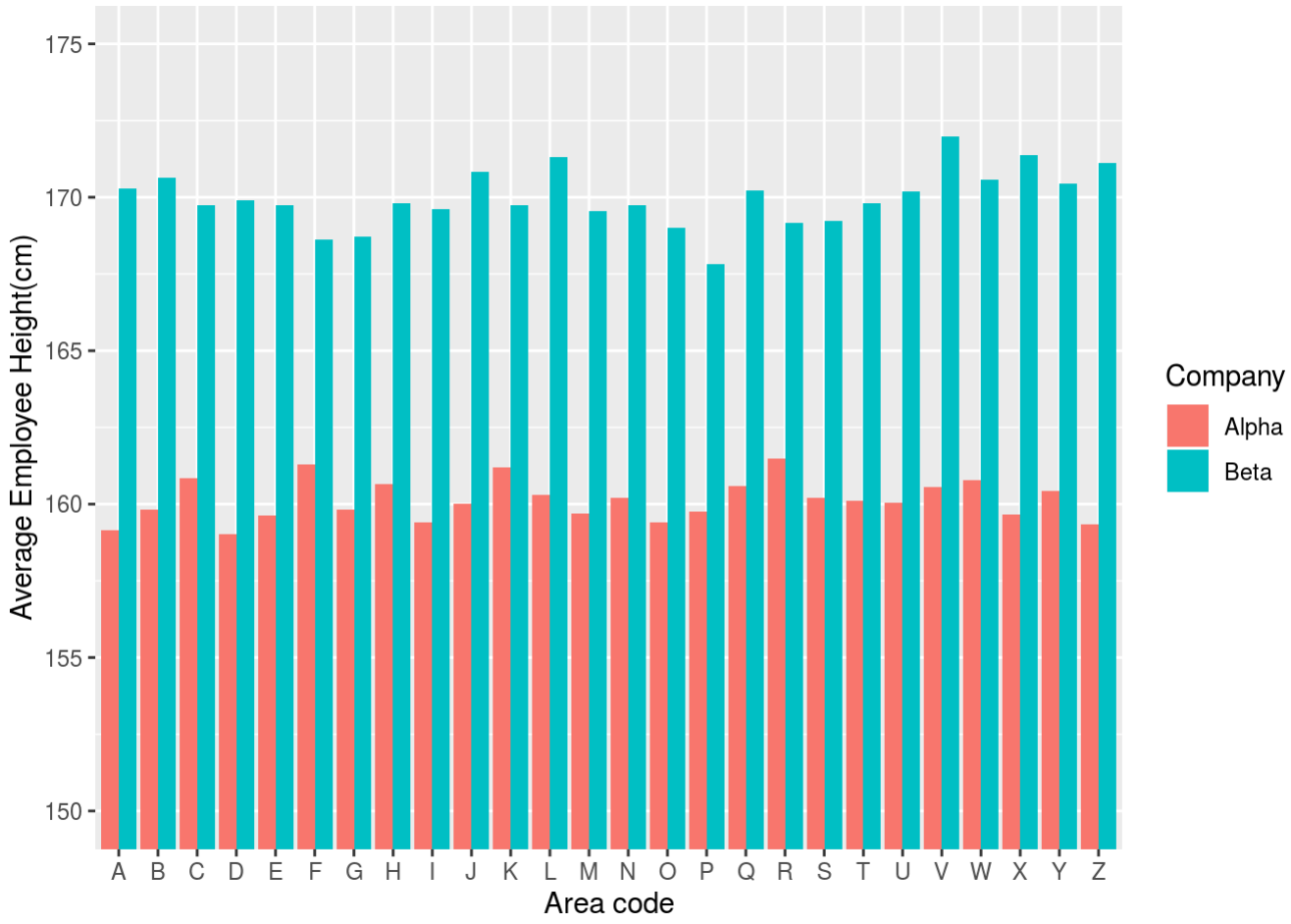
Problem 2

Still working on the previous data frame. For each area, summarize the average employee height of each company. Then plot a dodge bar chart visualizing area code versus the average of height, and mapping company as fill.

```
##### Summarising the data
new<-company_df%>%group_by(`Area code`,`Company`)%>%summarise(`Average Employee Height(cm)`=mean
(`Employee Height(cm)`))
new
```

```
## # A tibble: 52 x 3
## # Groups:   Area code [26]
##    `Area code` Company `Average Employee Height(cm)`
##    <fct>       <fct>                           <dbl>
##  1 A           Alpha                            159.
##  2 A           Beta                             170.
##  3 B           Alpha                            160.
##  4 B           Beta                             171.
##  5 C           Alpha                            161.
##  6 C           Beta                             170.
##  7 D           Alpha                            159.
##  8 D           Beta                             170.
##  9 E           Alpha                            160.
## 10 E           Beta                             170.
## # … with 42 more rows
```

```
##### Plotting the dodge bar graph
ggplot(new)+geom_col(aes(`Area code`,`Average Employee Height(cm)`,fill=`Company`),position = po
sition_dodge())+coord_cartesian(ylim=c(150,175))
```



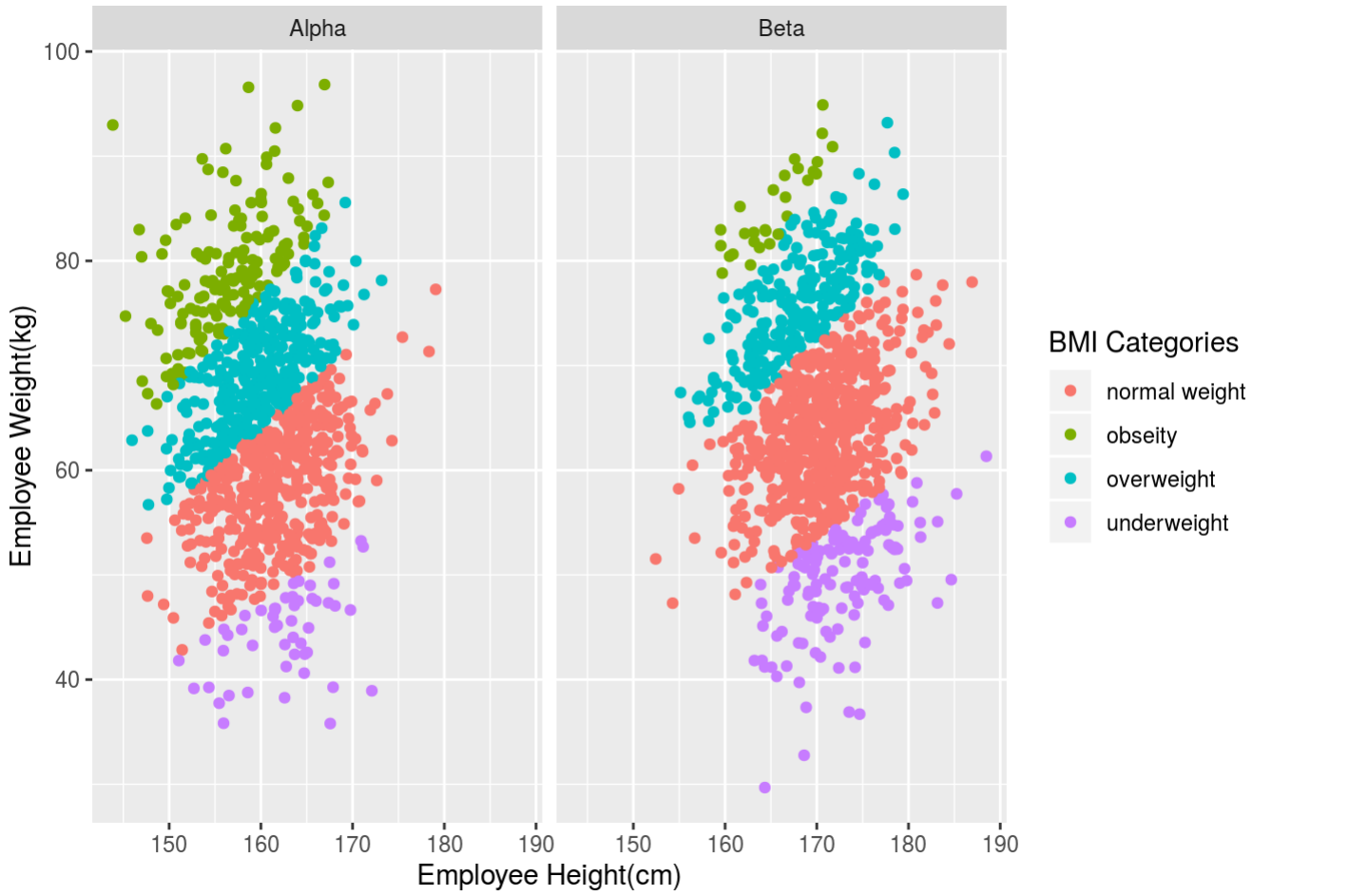
Problem 3

Insert THREE more columns into the previous data frame. • First column Employee Weight (kg) should be generated with 2000 random variables (mean = 65, sd = 10). • Second column “BMI” follows the formula: weight(kg)/[(height(cm)/100)ˆ2] • Third column BMI Categories contains 4 labels “underweight”, “normal weight”, “overweight”, and “obesity” associated with column “BMI” for each row. – When BMI <=18.5, “Underweight” – When 18.5< BMI<= 25, “Normal weight” – When 25< BMI <=30, “Overweight” – When BMI > 30, “Obesity” Then create a scatterplot visualizing Employee Height(cm) versus Employee Weight(kg), mapping BMI Categories as color, and facet this plot by Company.

```
##### Adding new columns to dataframe
company_df$`Employee Weight(kg)`<-rnorm(2000,mean=65,sd=10)

company_df$`BMI`<-(company_df$`Employee Weight(kg)`) / ((company_df$`Employee Height(cm)`/100)^2
)

company_df<-company_df%>%mutate(`BMI Categories`=case_when(`BMI`<=18.5~"underweight",`BMI`<=25 &
`BMI`>18.5~"normal weight",`BMI`<=30 & `BMI`>25~'overweight',`BMI`>30~'obseity'))

head(company_df)
```

```
##   Area code Company Employee Height(cm) Employee Weight(kg)      BMI
## 1         K   Alpha            162.9499            67.16248 25.29406
## 2         I   Alpha            157.7288            56.15715 22.57268
## 3         J   Alpha            158.9471            85.55954 33.86593
## 4         P   Alpha            164.6853            82.25004 30.32678
## 5         L   Alpha            157.0649            62.53163 25.34788
## 6         W   Alpha            162.4829            65.86176 24.94697
##   BMI Categories
## 1     overweight
## 2  normal weight
## 3        obseity
## 4        obseity
## 5     overweight
## 6  normal weight
```

```
##### Plotting scatter plot
ggplot(company_df)+geom_point(mapping=aes(x=`Employee Height(cm)`,y=`Employee Weight(kg)`,color=
`BMI Categories`))+facet_grid(~Company)
```



Section B

Section B uses National Health and Nutrition Examination Survey 2015-2016 Demographics Data from Centers for Disease Control and Prevention. Download NHANES 2015-2016 Demographics data (XPT file) from: https://wwwn.cdc.gov/nchs/nhanes/ (https://wwwn.cdc.gov/nchs/nhanes/) Search/DataPage.aspx?Component=Demographics&CycleBeginYear=2015

To read the data manual: https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/DEMO_I.htm (https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/DEMO_I.htm) The details and introduction for NHNES please click the link: https://youtu.be/GmnN2r5J0YA (https://youtu.be/GmnN2r5J0YA) Load package “haven”(one of the packages from “tidyverse”), and use read_xpt() to import the dataset to R.

Problem 1 Create a new data frame with the following columns: • The race information included only Mexican American, Other Hispanic, Non-Hispanic White, Non- Hispanic Black, and other race • Ratio/value of family income to the poverty line • Removing the above ratio’s decimals (e.g. 2.61 -> 2) and then make them as categorical data (“Annual family income value”): 0, 1, 2, 3, 4, and 5 • The proportion of each ethnic families among all families • The proportion of each ethnic families among all families at each annual family income value: 0, 1, 2, 3, 4, and 5

Then create a bar chart to visualize the annual family income value (x-axis) versus the proportion of Black families among all families at each annual family income value (y-axis). Include a subline whose y value should equal to the proportion of Black families among all families. Are Black families over- or under-represented in poverty? What else you notice about the chart? hints: When the annual family income value is 0, which means such family is in poverty

```
NHANE<-read_xpt('DEMO_I.XPT')
NHANE<-NHANE%>%drop_na(RIDRETH1)%>%drop_na(INDFMPIR)
head(NHANE)
```

```
## # A tibble: 6 x 47
##    SEQN SDDSRVYR RIDSTATR RIAGENDR RIDAGEYR RIDAGEMN RIDRETH1 RIDRETH3
##   <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
## 1 83732        9        2        1       62       NA        3        3
## 2 83733        9        2        1       53       NA        3        3
## 3 83734        9        2        1       78       NA        3        3
## 4 83735        9        2        2       56       NA        3        3
## 5 83736        9        2        2       42       NA        4        4
## 6 83737        9        2        2       72       NA        1        1
## # … with 39 more variables: RIDEXMON <dbl>, RIDEXAGM <dbl>,
## #   DMQMILIZ <dbl>, DMQADFC <dbl>, DMDBORN4 <dbl>, DMDCITZN <dbl>,
## #   DMDYRSUS <dbl>, DMDEDUC3 <dbl>, DMDEDUC2 <dbl>, DMDMARTL <dbl>,
## #   RIDEXPRG <dbl>, SIALANG <dbl>, SIAPROXY <dbl>, SIAINTRP <dbl>,
## #   FIALANG <dbl>, FIAPROXY <dbl>, FIAINTRP <dbl>, MIALANG <dbl>,
## #   MIAPROXY <dbl>, MIAINTRP <dbl>, AIALANGA <dbl>, DMDHHSIZ <dbl>,
## #   DMDFMSIZ <dbl>, DMDHHSZA <dbl>, DMDHHSZB <dbl>, DMDHHSZE <dbl>,
## #   DMDHRGND <dbl>, DMDHRAGE <dbl>, DMDHRBR4 <dbl>, DMDHREDU <dbl>,
## #   DMDHRMAR <dbl>, DMDHSEDU <dbl>, WTINT2YR <dbl>, WTMEC2YR <dbl>,
## #   SDMVPSU <dbl>, SDMVSTRA <dbl>, INDHHIN2 <dbl>, INDFMIN2 <dbl>,
## #   INDFMPIR <dbl>
```

```
new_NHANE<-NHANE%>%mutate(Race=case_when(RIDRETH1==1~'Mexican American',RIDRETH1==2~'Other Hispa
nic',RIDRETH1==3~'Non-Hispanic',RIDRETH1==4~'Non Hispanic Black',RIDRETH1==5~'other Races'))%>%r
ename('Ratio of family income to poverty'=`INDFMPIR`)
```

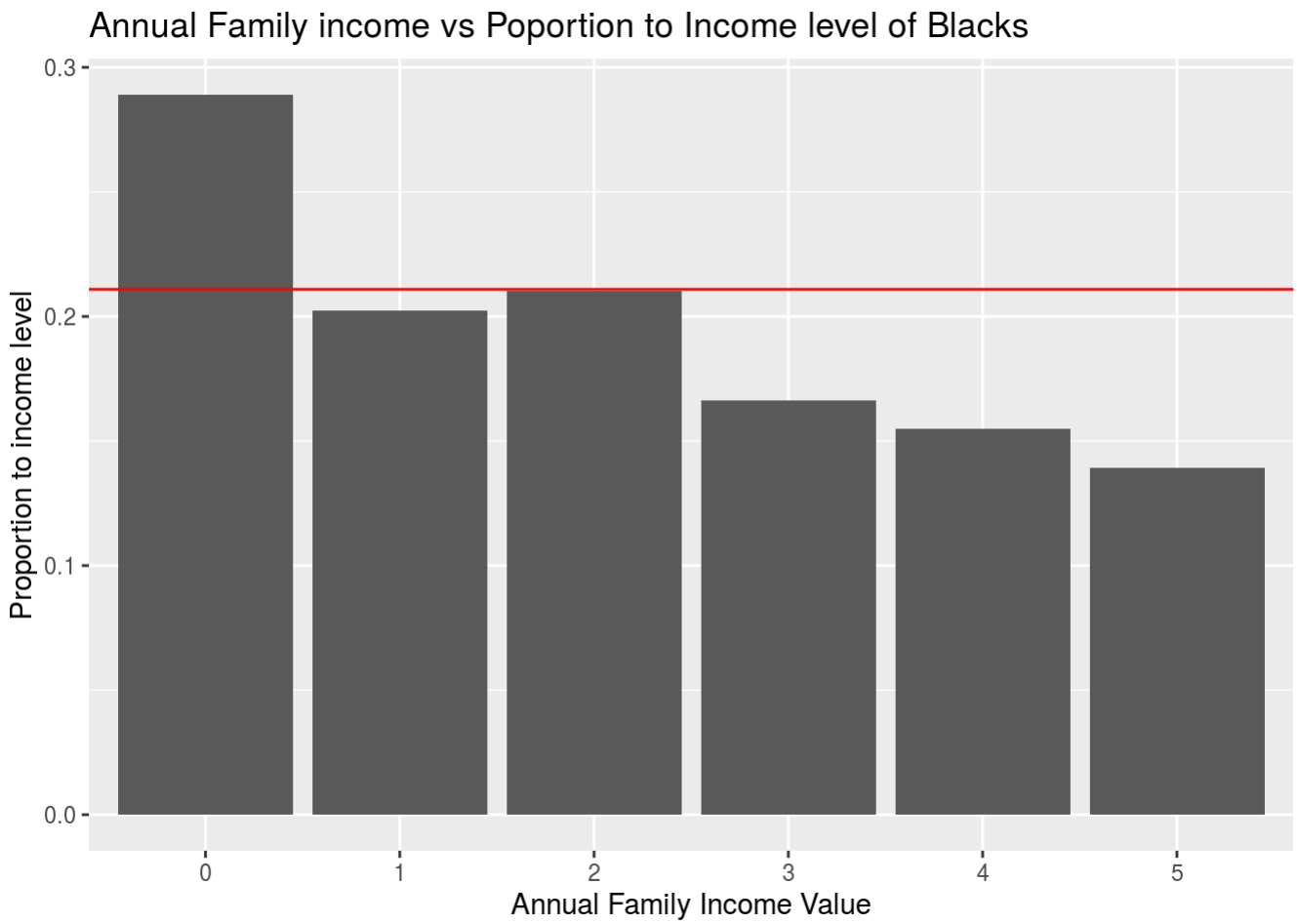
```
new_NHANE<-new_NHANE%>%select(`Ratio of family income to poverty`,Race)
new_NHANE$`Annual Family Income Value`<-as.character(floor(new_NHANE$`Ratio of family income to
poverty`))
new_NHANE<-new_NHANE[c(2,1,3)]
head(new_NHANE)
```

```
## # A tibble: 6 x 3
##   Race             `Ratio of family income to po… `Annual Family Income Va…
##   <chr>                                     <dbl> <chr>
## 1 Non-Hispanic                               4.39 4
## 2 Non-Hispanic                               1.32 1
## 3 Non-Hispanic                               1.51 1
## 4 Non-Hispanic                               5    5
## 5 Non Hispanic Bl…                           1.23 1
## 6 Mexican American                           2.82 2
```

```
prop <-new_NHANE %>% group_by(Race) %>% summarise(total = n()) %>% mutate(Proportion = total / s
um(total))
new_NHANE<-new_NHANE%>%inner_join(prop, by='Race')

Prop2 <-new_NHANE%>% group_by(Race, `Annual Family Income Value`,)%>%summarise(total = n())%>% m
utate('Proportion to income level' = total/table(new_NHANE$`Annual Family Income Value`))

new <-new_NHANE%>%inner_join(Prop2,by = c("Race", "Annual Family Income Value"))
head(new)
```

```
## # A tibble: 6 x 7
##   Race  `Ratio of famil… `Annual Family … total.x Proportion total.y
##   <chr>            <dbl> <chr>              <int>      <dbl>   <int>
## 1 Non-…             4.39 4                   2877      0.323     262
## 2 Non-…             1.32 1                   2877      0.323     783
## 3 Non-…             1.51 1                   2877      0.323     783
## 4 Non-…             5    5                   2877      0.323     586
## 5 Non …             1.23 1                   1881      0.211     501
## 6 Mexi…             2.82 2                   1665      0.187     265
## # … with 1 more variable: `Proportion to income level` <dbl>
```

```
##### Creating a bar chart
black<-Prop2%>%filter(Race=='Non Hispanic Black')

ggplot(black,aes(x=`Annual Family Income Value`,y=`Proportion to income level`))+geom_col()+geom
_hline(aes(yintercept =prop$Proportion[prop$Race=='Non Hispanic Black'] ), colour = "red")+ggtit
le('Annual Family income vs Poportion to Income level of Blacks')
```



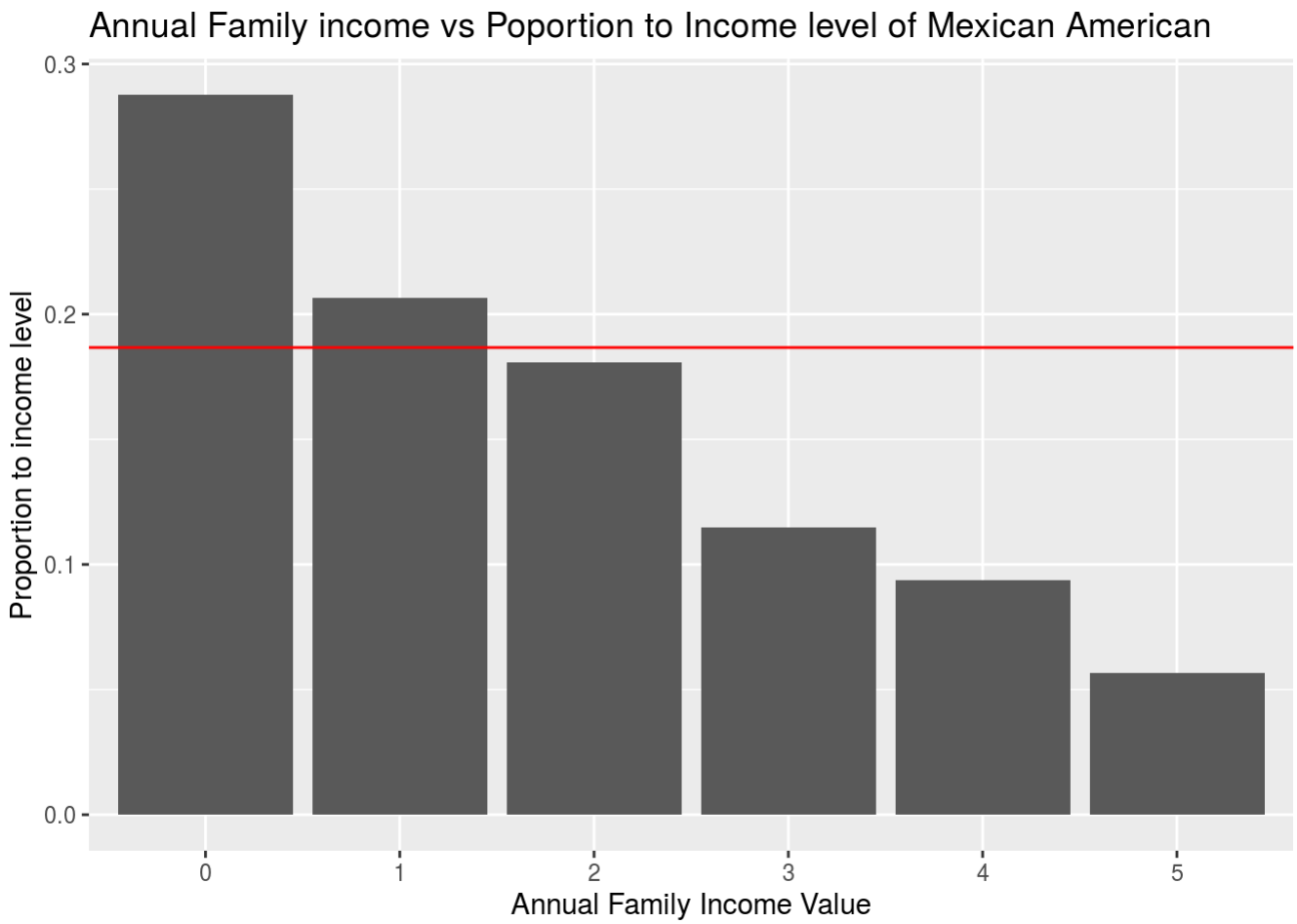
Interpretation

The x-axis gives the Annual family value vs proportion to income level. The subline gives the proportion of non hispanic black to overall ethinic family. The Annual income of proportion of Blacks are fluctuated. The subline indicates the median proportion and annual income of zero indicates thata] the blacks are over represented in poverty. The annual income from 1 to 5 indicates that the blacks are under represented to the overall proportion of blacks.

Problem 2

Still working on the above data frame. Then create a bar chart to visualize the annual family income value (x-axis) versus the proportion of Mexican American families among all families at each annual family income value (y-axis). Include a subline whose y value should equal to the proportion of Mexican American families among all families. Are Mexican American families over- or under-represented in poverty? What else you notice about the chart?

```
##### Plotting the bar graph
mexican<-Prop2%>%filter(Race=='Mexican American')

ggplot(mexican)+geom_col(mapping=aes(x=`Annual Family Income Value`,y=`Proportion to income leve
l`))+geom_hline(aes(yintercept =prop$Proportion[prop$Race=='Mexican American'] ), colour = "red"
)+ggtitle('Annual Family income vs Poportion to Income level of Mexican American')
```

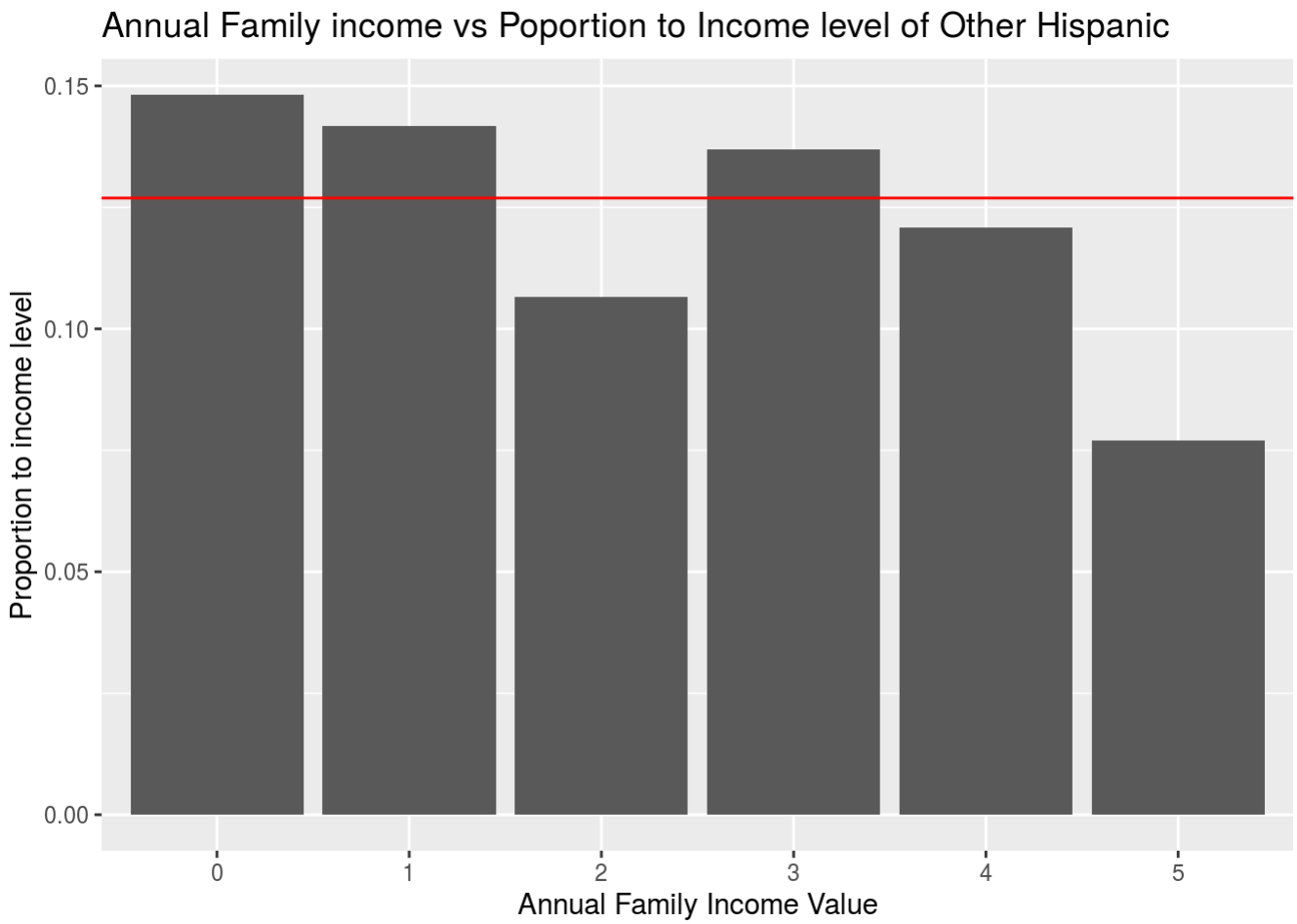


interpretation:

The x-axis gives the Annual family value of Mexican American vs proportion to income level. The subline gives the proportion of Mexican American to overall ethinic family. The Annual income of Mexican American are decreasing and almost steeped at the end. The zero annual income represents that the family are in poverty and the subline indicates the median scale of Mexican American which proofs that the Mexican American are over represented in poverty. The income level of 1 is also overrepresented when compared to the median of proportion.The income level from 2 to 5 are under represented when compared to the medium of proportion.

Problem 3 Still working on the above data frame. Select other hispanic families for observation. Then create a bar chart to visualize the annual family income value (x-axis) versus the proportion of other hispanic families among all families at each annual family income value (y-axis). Include a subline whose y value should equal to the proportion of other hispanic families among all families. Are other hispanic families over- or under-represented in poverty? What else you notice about the chart?

```
##### Plotting the bar graph
hispanic<-Prop2%>%filter(Race=='Other Hispanic')

ggplot(hispanic)+geom_col(mapping=aes(x=`Annual Family Income Value`,y=`Proportion to income lev
el`))+geom_hline(aes(yintercept =prop$Proportion[prop$Race=='Other Hispanic'] ), colour = "red")
+ggtitle('Annual Family income vs Poportion to Income level of Other Hispanic')
```



Interpretation: The x-axis gives the Annual family value of Other Hispanic vs proportion to income level. The subline gives the proportion of Other Hispanic to overall ethinic family. The Annual income of Other Hispanic is fluctuated. The proportion to income level of 0,1,3 all families are high indicating that they are over represented. The other hispanic family are overpresented in the consideration on poverty, which shows zero is higher than the average proportion of other hispanic.While the Annual income level of 2,4 and 5 are under represented.