Assignment2

PartA Question1

```
library(tidyverse)
(df1 <- as.tibble(read_csv("silicon.csv")))
```

```
## # A tibble: 3,960 x 6
##    company  year   race gender                        job_category count
##      <chr> <int>  <chr>  <chr>                               <chr> <chr>
##  1 23andMe  2016 Latino   male Executive/Senior officials & Mgrs     0
##  2 23andMe  2016 Latino   male        First/Mid officials & Mgrs     1
##  3 23andMe  2016 Latino   male                     Professionals     7
##  4 23andMe  2016 Latino   male                       Technicians     0
##  5 23andMe  2016 Latino   male                     Sales workers     0
##  6 23andMe  2016 Latino   male            Administrative support     0
##  7 23andMe  2016 Latino   male                     Craft workers     0
##  8 23andMe  2016 Latino   male                        operatives     0
##  9 23andMe  2016 Latino   male               laborers and helpers     0
## 10 23andMe  2016 Latino   male                   Service workers     0
## # ... with 3,950 more rows
```

The data investigates the demographics for 23 Silicon Valley tech companies.

Question2

```
df2 <- df1%>%
  group_by(company)%>%
  filter(race=="Asian")%>%
  summarise(Asian=sum(as.numeric(count),na.rm=T)/2)
```

```
## Warning in eval(substitute(expr), envir, enclos): NAs introduced by
## coercion

## Warning in eval(substitute(expr), envir, enclos): NAs introduced by
## coercion

## Warning in eval(substitute(expr), envir, enclos): NAs introduced by
## coercion

## Warning in eval(substitute(expr), envir, enclos): NAs introduced by
## coercion
```

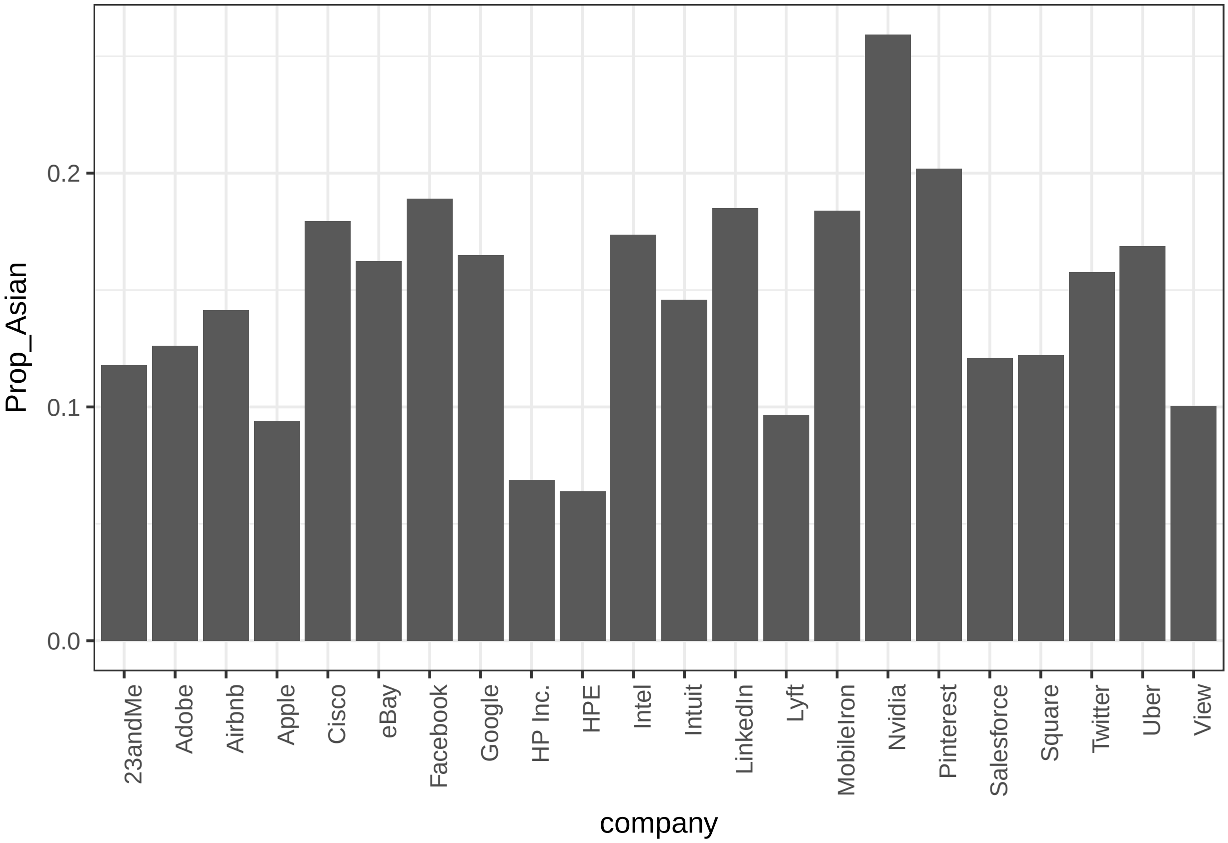
```
df3 <- df1%>%
  group_by(company)%>%
  summarise(Total=sum(as.numeric(count),na.rm=T)/2)
```

```
## Warning in eval(substitute(expr), envir, enclos): NAs introduced by
## coercion

## Warning in eval(substitute(expr), envir, enclos): NAs introduced by
## coercion

## Warning in eval(substitute(expr), envir, enclos): NAs introduced by
## coercion
## Warning in eval(substitute(expr), envir, enclos): NAs introduced by
## coercion
```

```
(df4 <- inner_join(df2,df3)%>%mutate(Prop_Asian=Asian/Total))
```

```
## # A tibble: 22 x 4
##     company   Asian  Total Prop_Asian
##       <chr>   <dbl>  <dbl>      <dbl>
##  1  23andMe    70.0    594 0.11784512
##  2    Adobe  2637.0  20905 0.12614207
##  3   Airbnb   739.5   5235 0.14126074
##  4    Apple 21329.5 226878 0.09401308
##  5    Cisco 19974.0 111366 0.17935456
##  6     eBay  3910.5  24082 0.16238269
##  7 Facebook  5847.0  30928 0.18905199
##  8   Google 21779.5 132191 0.16475781
##  9  HP Inc.  6830.0  99377 0.06872818
## 10      HPE  6634.0 103978 0.06380196
## # ... with 12 more rows
```

```
df4%>%
  arrange(Prop_Asian)%>%
  ggplot(aes(x=company,y=Prop_Asian))+
  theme_bw()+geom_bar(stat="identity")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

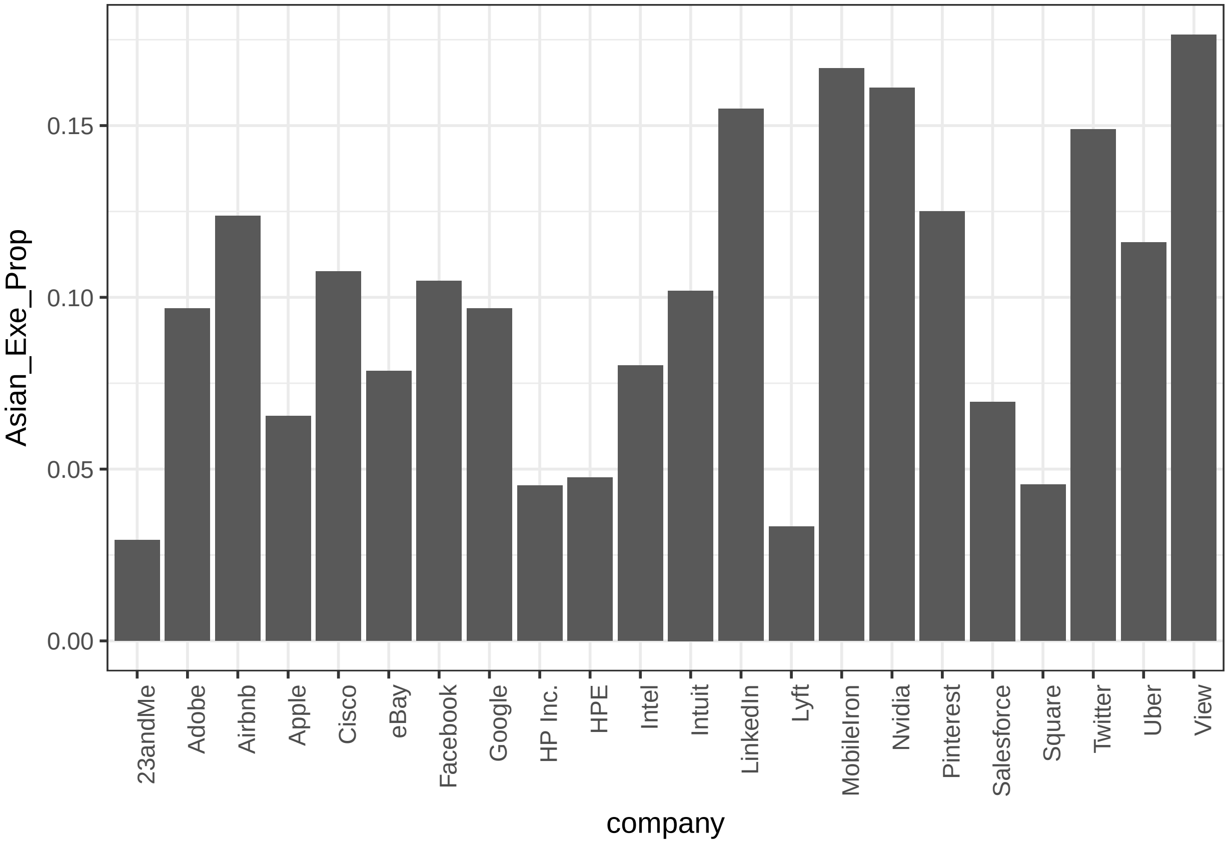


From the graph, we can know that the proportion of Asian ranges from 6.38% to 25.9%. HPE has the lowest proportion of Asian employees, and Navidia has the greatest proportion of Asian employees.

```
df5 <- df1%>%
  group_by(company)%>%
  filter(job_category=="Executive/Senior officials & Mgrs")%>%
  summarise(Executive=sum(as.numeric(count),na.rm=T))

df6 <- df1%>%
  group_by(company)%>%
  filter(race=="Asian")%>%
  filter(job_category=="Executive/Senior officials & Mgrs")%>%
  summarise(Asian_Exe=sum(as.numeric(count),na.rm=T))

df7 <- df5%>%inner_join(df6)%>%mutate(Asian_Exe_Prop=Asian_Exe/Executive)
```

```
## Joining, by = "company"
```

```
df7%>%ggplot(aes(x=company,y=Asian_Exe_Prop))+
  theme_bw()+
  geom_bar(stat="identity")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

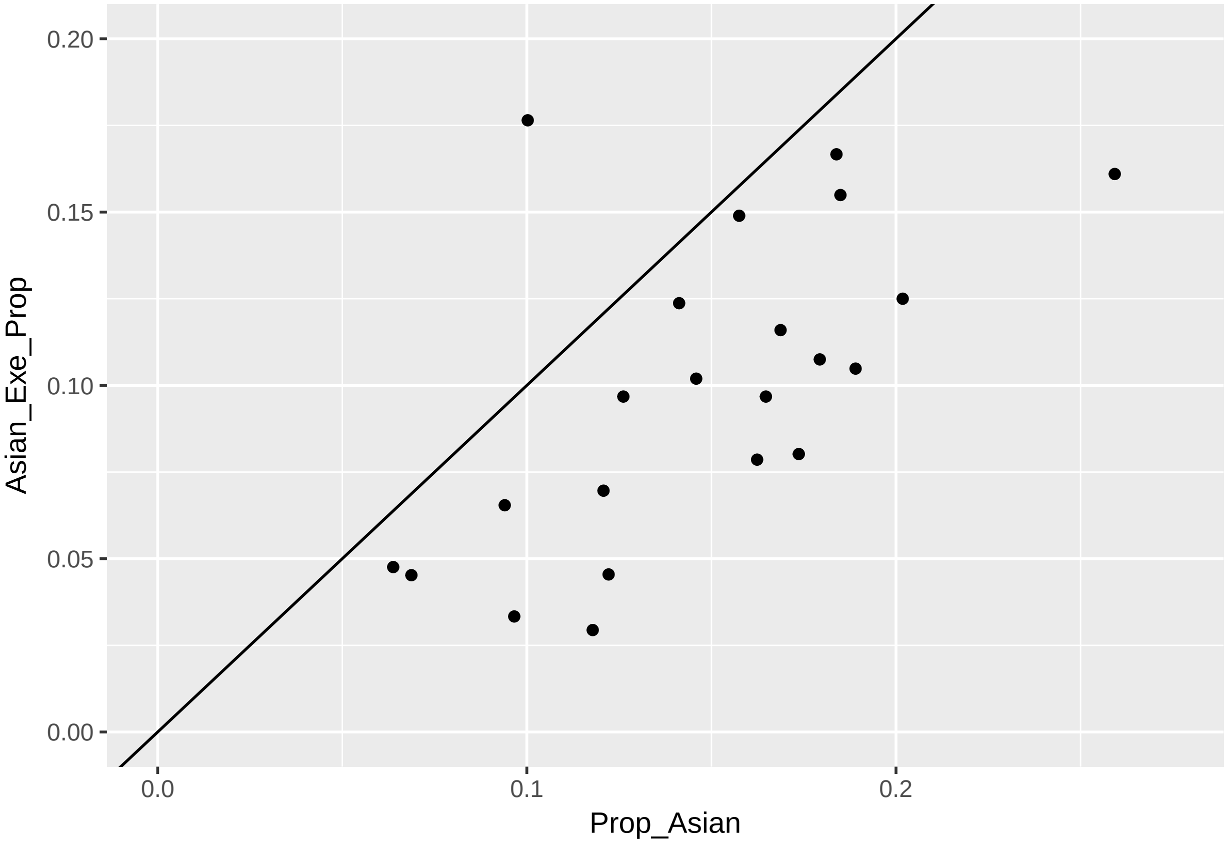


The graph shows the proportion of executive who is Asian in each company, which ranges from 2.9% to 17.6%. View has the highest proportion of executive who is Asian, while 23andMe has the lowest proportion of asian executive.

```
df8 <- df7%>%inner_join(df4)
```

```
## Joining, by = "company"
```

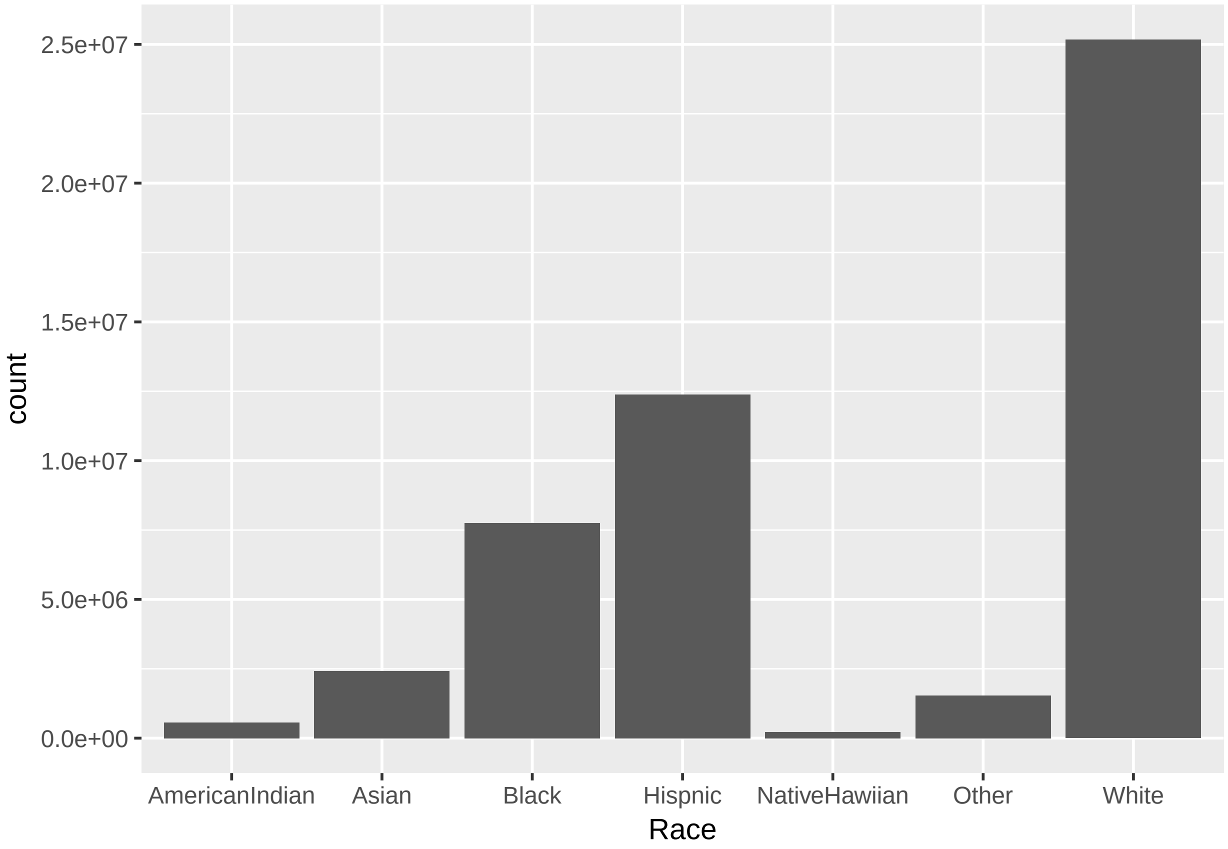
```
df8%>%group_by(company)%>%
  ggplot(aes(x=Prop_Asian,y=Asian_Exe_Prop))+
  geom_point()+geom_abline(slope = 1,intercept = 0.0)+
  coord_cartesian(ylim=c(0, 0.2),xlim = c(0,0.275))
```



This graph combines the proportion of asian and proportion of asian executive for each company. A straight line x=y is added to indicate a under-representation of Asian who take leadership in a company.

Part B Question 3

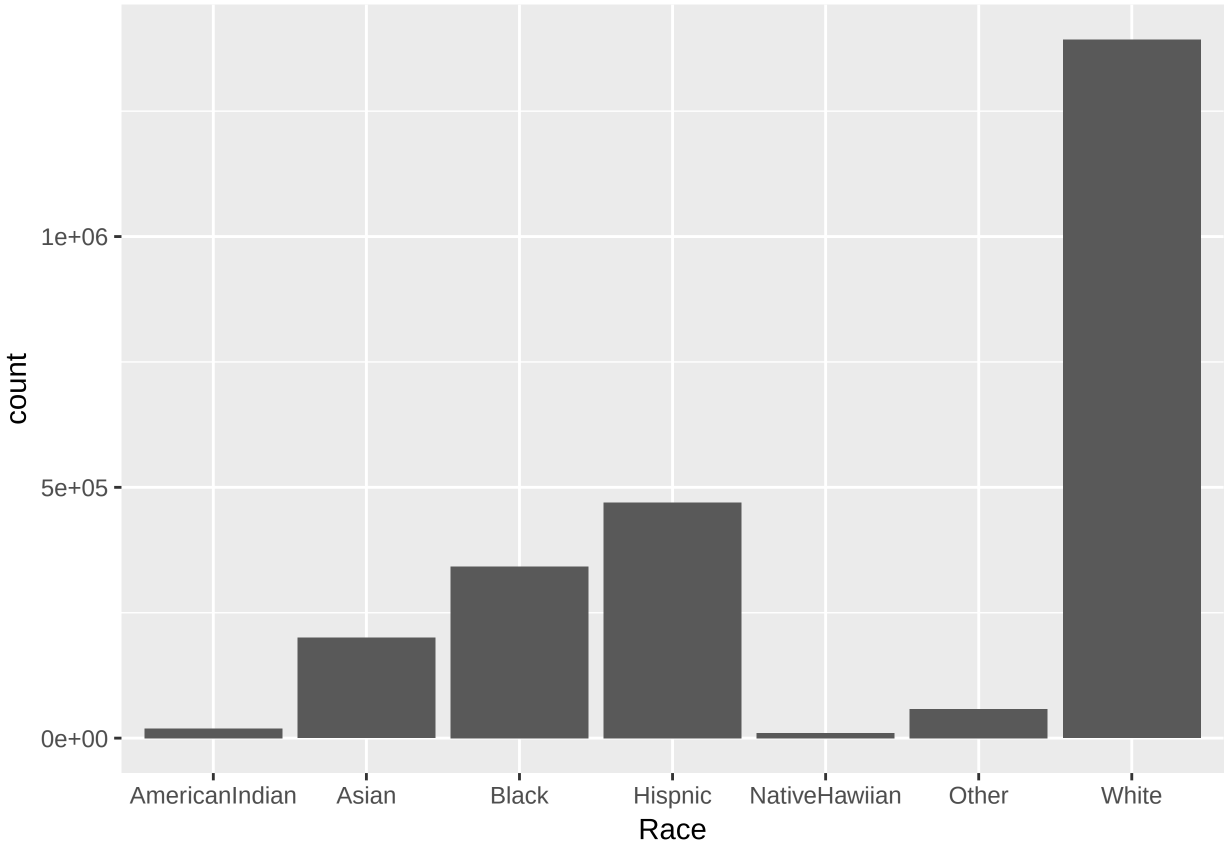
```
df1_b<- read_csv("CRDC2013_14_SCH.csv",na=c("-2","-5","-9"))
df2_b <- df1_b%>%transmute(Hispnic=SCH_ENR_HI_M+SCH_ENR_HI_F,
                           AmericanIndian=SCH_ENR_AM_M+SCH_ENR_AM_F,
                           Asian=SCH_ENR_AS_M+SCH_ENR_AS_F,
                           NativeHawiian=SCH_ENR_HP_M+SCH_ENR_HP_F,
                           Black=SCH_ENR_BL_M+SCH_ENR_BL_F,
                           White=SCH_ENR_WH_M+SCH_ENR_WH_F,
                           Other=SCH_ENR_TR_M+SCH_ENR_TR_F) %>%
  summarize(Hispnic=sum(Hispnic,na.rm=TRUE),
            AmericanIndian=sum(AmericanIndian,na.rm=TRUE),
            Asian=sum(Asian,na.rm=TRUE),
            NativeHawiian=sum(NativeHawiian,na.rm=TRUE),
            Black=sum(Black,na.rm=TRUE),
            White=sum(White,na.rm=T),
            Other=sum(Other,na.rm=T)
  )
gather(df2_b,Hispnic,AmericanIndian,Asian,
       NativeHawiian,Black,White,Other,
       key= "Race", value= "count") %>%
  ggplot(aes(x=Race,y=count))+geom_col()
```



The graph shows the race against the number of population of each race.

Question4

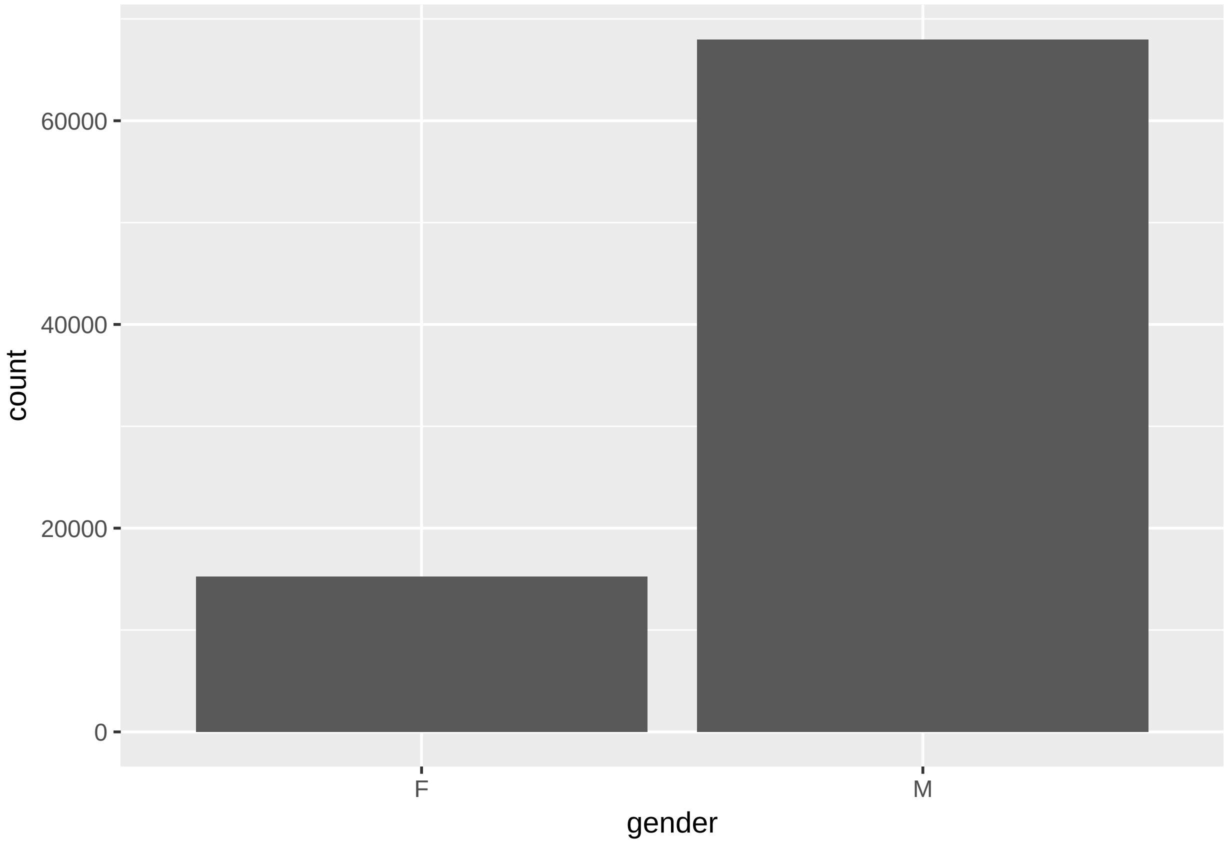
```
df3_b <- df1_b%>%
  transmute(Hispnic=SCH_MATHENR_ADVM_HI_M+SCH_MATHENR_ADVM_HI_F,
     AmericanIndian=SCH_MATHENR_ADVM_AM_M+SCH_MATHENR_ADVM_AM_F,
    Asian=SCH_MATHENR_ADVM_AS_M+SCH_MATHENR_ADVM_AS_F,
    NativeHawiian=SCH_MATHENR_ADVM_HP_M+SCH_MATHENR_ADVM_HP_F,
   Black=SCH_MATHENR_ADVM_BL_M+SCH_MATHENR_ADVM_BL_F,
   White=SCH_MATHENR_ADVM_WH_M+SCH_MATHENR_ADVM_WH_F,
   Other=SCH_MATHENR_ADVM_TR_M+SCH_MATHENR_ADVM_TR_M)%>%
  summarize(Hispnic=sum(Hispnic,na.rm=TRUE),
            AmericanIndian=sum(AmericanIndian,na.rm=TRUE),
            Asian=sum(Asian,na.rm=TRUE),
            NativeHawiian=sum(NativeHawiian,na.rm=TRUE),
            Black=sum(Black,na.rm=TRUE),
            White=sum(White,na.rm=T),
            Other=sum(Other,na.rm=T))
gather(df3_b,Hispnic,AmericanIndian,Asian,
       NativeHawiian,Black,White,Other,
       key= "Race", value= "count") %>%
  ggplot(aes(x=Race,y=count))+geom_col()
```



The graph from Question 3 has a similar shape with the graph from Question 4, where the greatest number of students who take advanced mathematicis is white, and the smallest number of students who take davabced mathematicis is Native Hawiian.

PartC Question5

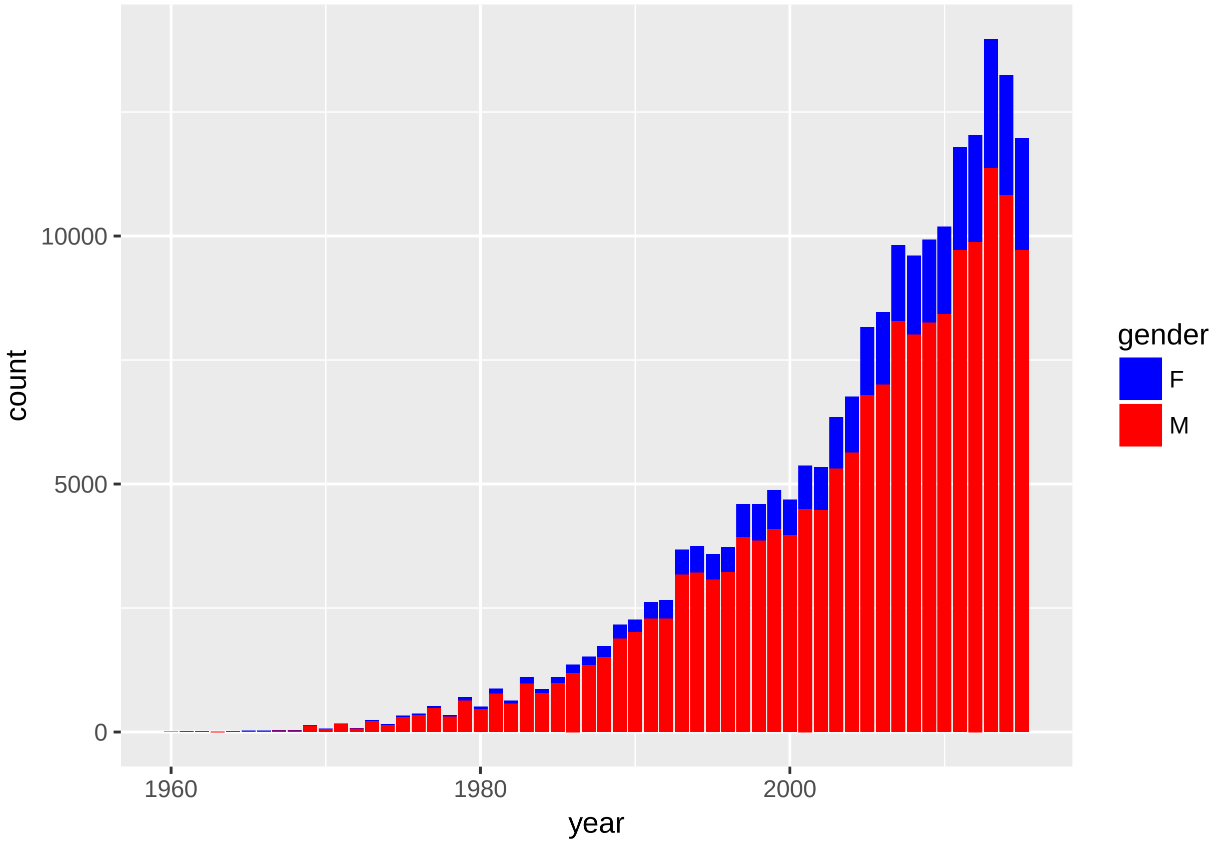
```
library(DBI)
library(RMySQL)
library(dbplyr)
con <- dbConnect(MySQL(),user="root", password="Hzy19940928.",dbname="dblp")
df1_c<- dbReadTable(con, 'general')
df2_c <- dbReadTable(con,'authors')
df3_c <- left_join(df1_c,df2_c)%>%filter(prob>=0.99 & prob<=1.00)
df3_c%>%group_by(gender)%>%
  summarise(count = n_distinct(name)) %>%
  collect() %>%
  ggplot(aes(x=gender,y=count))+
  geom_bar(stat="identity")
```



The graph shows the the number of distinct male and female authors in the dataset.

Question6

```
df3_c%>%
  group_by(gender,year)%>%
  summarise(count = n_distinct(name))%>%
  ggplot(aes(x=year,y=count,fill=gender)) +geom_bar(stat="identity",position = "stack")+
  scale_fill_manual(values=c("blue","red"))
```

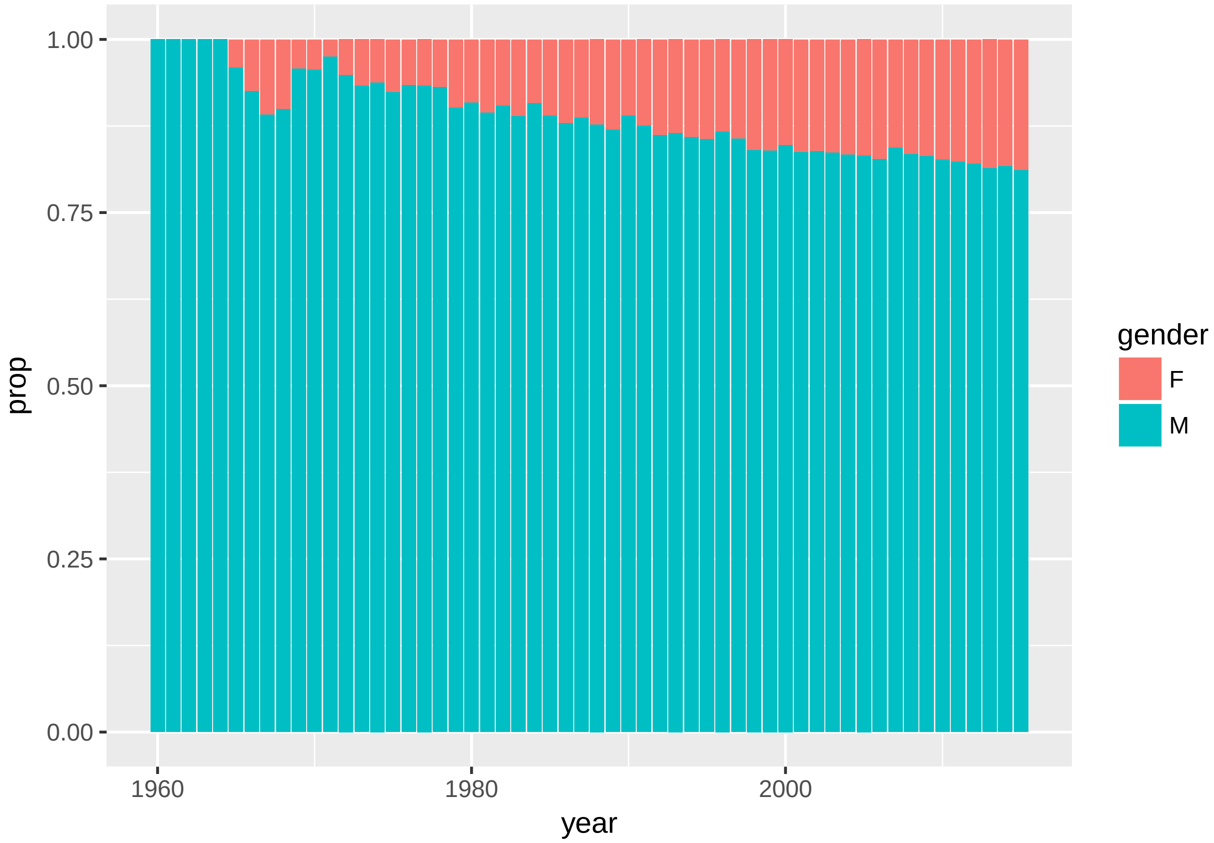


The graph is a stacked bar plot showing the number of distinct male and female authors published each year.

Question7

```
df4_c_1<- df3_c%>%group_by(year)%>%summarise(Total=n_distinct(name))
df4_c_2 <- df3_c%>%group_by(gender,year)%>%
  summarise(Total_gender=n_distinct(name))

left_join(df4_c_2,df4_c_1)%>%
  group_by(year,gender)%>%
  summarise(prop=Total_gender/Total,na.rm=TRUE)%>%
  ggplot(aes(x=year,y=prop,fill=gender))+
  geom_bar(stat="identity",position="stack")
```

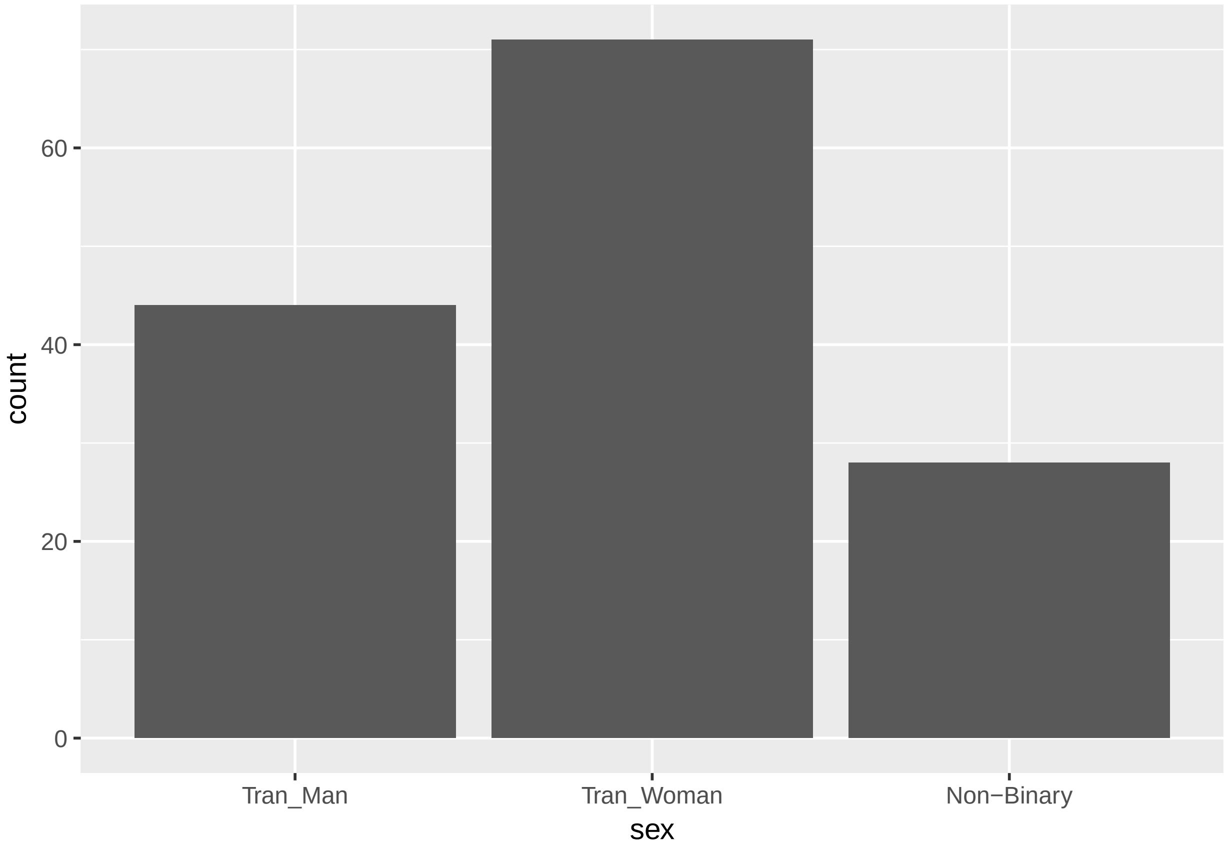
```
## Joining, by = "year"
```



The graph is a stacked bar plot showing the proportions of distinct male and female authors published each year.

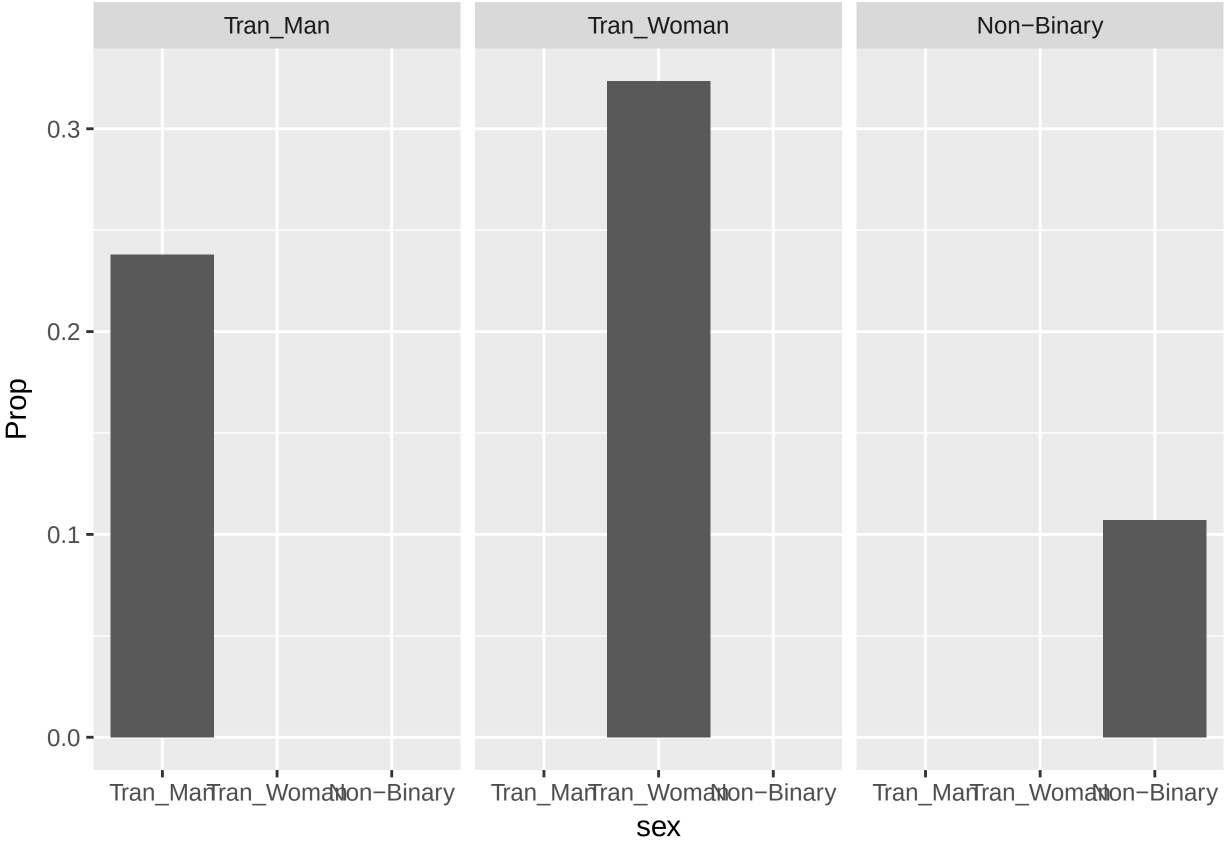
Question8

```
load(file="31721-0001-Data.rda")
df1_d=da31721.0001
df_d_TW <- df1_d%>%
  filter(is.na(Q6)!=T)%>%
  filter(Q6=="(2) Woman")%>%
  filter(Q5=="(1) Male")
df_d_M <- df1_d%>%
  filter(is.na(Q6)!=T)%>%
  filter(Q5=="(2) Female")%>%
  filter(Q6=="(1) Man")
df_d_NB <- df1_d%>%
  filter(is.na(Q6)!=T)%>%
  filter(Q6=="(4) Androgynous" | Q6=="(6) Gender Queer")
df_d_8 <- rbind(df_d_TW,df_d_M,df_d_NB)
df_d_8_1 <- df_d_8%>%transmute(sex=Q6,Denied=Q84,Fried=Q86)%>%
  mutate(sex=recode(sex,
                    "(4) Androgynous"="Non-Binary",
                    "(6) Gender Queer"="Non-Binary",
                    "(1) Man"="Tran_Man",
                    "(2) Woman"="Tran_Woman" ))
df_d_8_1%>%ggplot(aes(x=sex))+geom_bar()
```



The graph shows the number of participants of each of the genders.

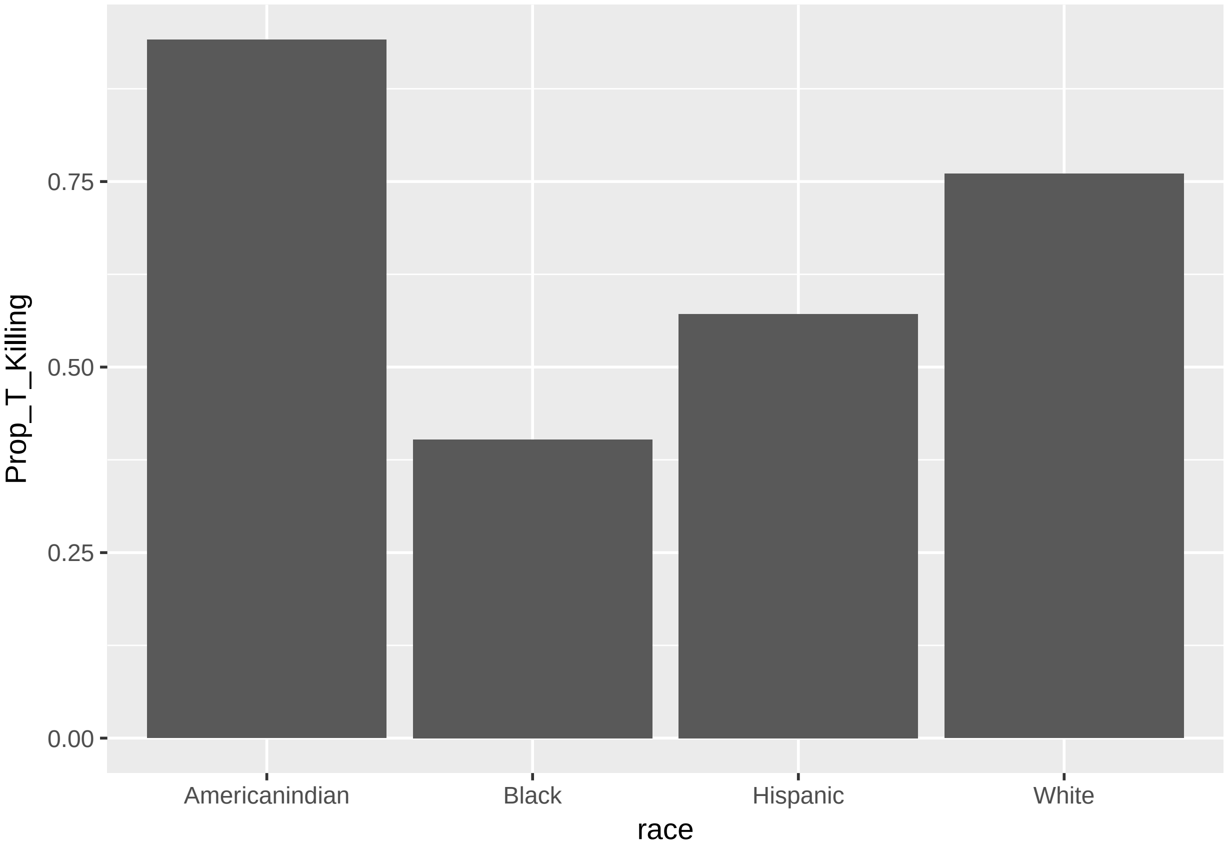
```
df_d_8_1%>%group_by(sex)%>%
  summarise(Prop=mean(Denied=="(1) Yes"|Fried=="(1) Yes",na.rm=T))%>%
ggplot(aes(x=sex,y=Prop))+geom_bar(stat="identity") +
  facet_wrap(~ sex)
```



The graph shows the proportion of participants who have been fired or denied a job due to their transgender status and/or gender expression.

Question9

```
df_d_9 <- da31721.0001%>%
  transmute(T_K=Q131,
            Black=D9_1,
            White=D9_2,
            Hispanic=D9_3,
            Americanindian=D9_4)%>%
  gather(key="race", value="isit",
         Black, White, Hispanic,Americanindian) %>%
  filter(isit=="(1) Selected") %>%
  select(-isit)%>%
  filter(T_K!="NA")

df_d_9%>%
  group_by(race)%>%
  summarise(Prop_T_Killing= sum(T_K=="(1) Yes",na.rm=T)/n())%>%
  ggplot(aes(x=race,y=Prop_T_Killing))+geom_bar(stat="identity")
```



```
da31721.0001%>%
  transmute(HAS=Q133)%>%
  summarise(Prop_HAS= sum(HAS=="(1) Yes",na.rm=T)/n())
```

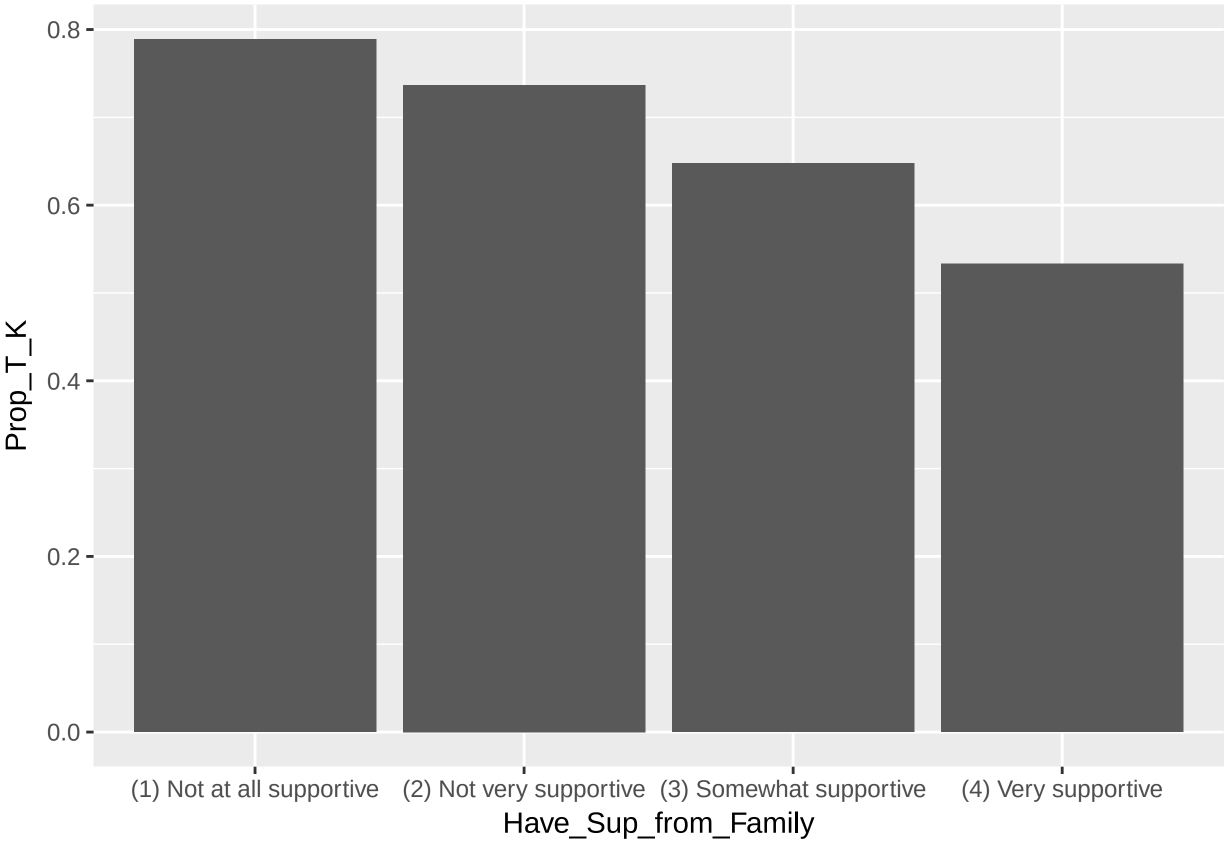
```
##    Prop_HAS
## 1 0.2542857
```

The graph shows the proportions of participants who have thought about killing themselves for African American, Caucasian, Hispanic/Latinx, and Native American demographics.

And the table determines total proportion of participants who have attempted suicide in the Virginia THIS survey, which is 25.43%. The calculated proportion is a lower than 41%.

Question10

```
df_d_10 <- da31721.0001%>%
  transmute(Have_Sup_from_Family=Q119,
            T_K=Q131)%>%
  filter(Have_Sup_from_Family!="(5) Not applicable to me")%>%
  filter(Have_Sup_from_Family!="NA")
df_d_10%>%
  group_by(Have_Sup_from_Family)%>%
  summarise(Prop_T_K=sum(T_K=="(1) Yes",na.rm=T)/n())%>%
  ggplot(aes(x=Have_Sup_from_Family,y= Prop_T_K))+
  geom_bar(stat="identity")
```



The graph shows the proportions of participants who have thought about killing themselves for each level of familial support. It indicates that, the support from the family is able to reduce the risk of suicide. The more support from the family, the less likely one will think of suicide.