hw3-Rohit-Thakur

Rohit Thakur

2/9/2020

R Markdown

PART A

PROBLEM 1

```
library(readr)
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

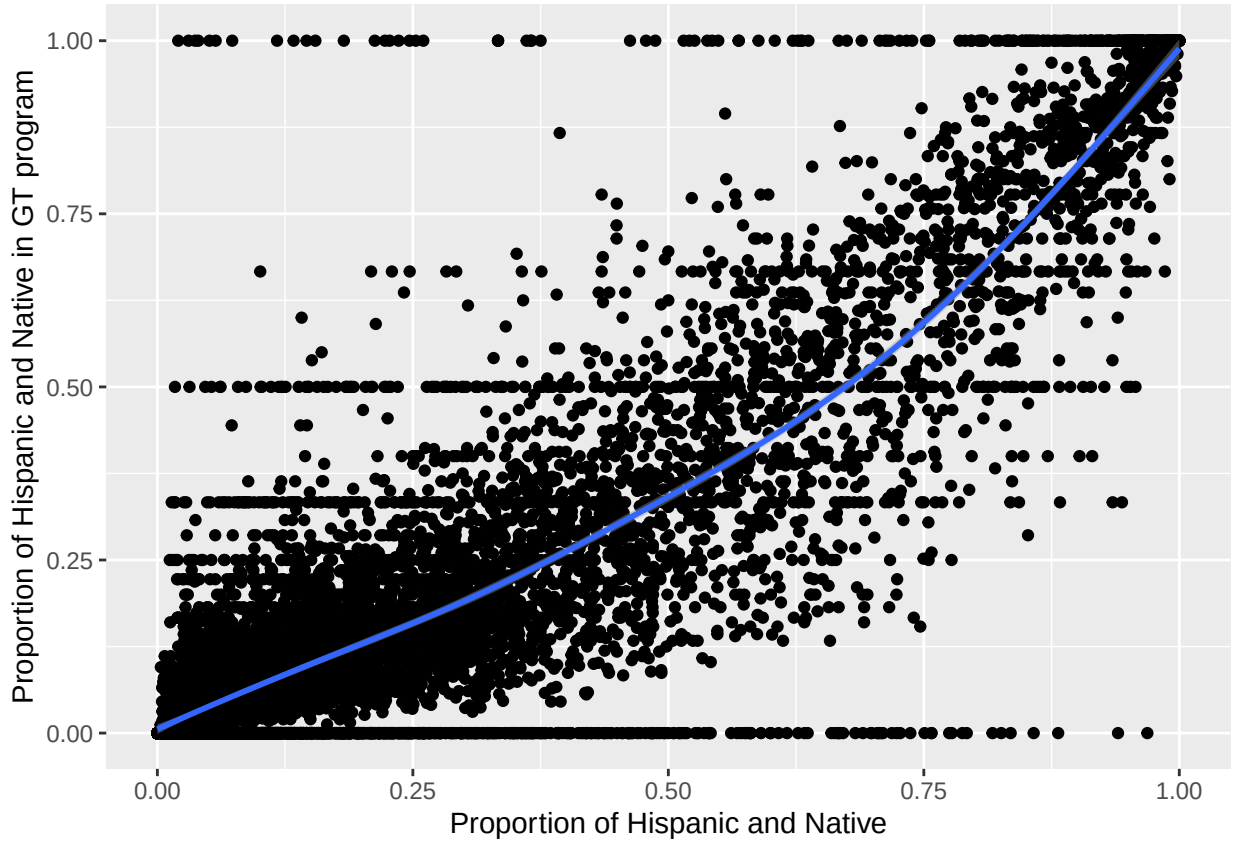
```
library(ggplot2)
school<-read.csv("D:/Spring 20 Sem 2/DMP/2015-16-crdc-data (1)/Data Files and Layouts/CRDC 2015-16 Scho
#Filtering for schools containing GT program
school<-filter(school,school$SCH_GT_IND == "Yes")
school1<-school
#Total Enrollment
school1$Tot_enr<-school1$TOT_ENR_M + school1$TOT_ENR_F
#Total students in gt program
school1$Tot_stud_gt<-school$TOT_GTENR_M + school$TOT_GTENR_F
#total hispanic and native americans in gt program
school1$Tot_HN_GT<-school1$SCH_GTENR_HI_M +
  school1$SCH_GTENR_HI_F +
  school1$SCH_GTENR_AM_M + school1$SCH_GTENR_AM_F
#total hispanic and native americans in school
school1$TOT_HN<-school1$SCH_ENR_HI_M +
  school1$SCH_ENR_HI_F +
  school1$SCH_ENR_AM_F + school1$SCH_ENR_AM_M
#Proportion of hispanic or native americans amongst everyone in the school
school1$PROP_HN<-(school1$TOT_HN)/(school1$Tot_enr)
#Proportion of hispanic or native americans in gt program amongst evryone in the gt program
school1$PROP_HN_GT<-(school1$Tot_HN_GT)/(school1$Tot_stud_gt)
schgt<-select(school1,Tot_enr,Tot_stud_gt,Tot_HN_GT,TOT_HN,PROP_HN,PROP_HN_GT)
#New dataframe with required columns
schgt%>%sample_n(10000)%>%ggplot(aes(x=PROP_HN,y=PROP_HN_GT))+
  geom_point()+
```

```
  geom_smooth()+
  xlab("Proportion of Hispanic and Native") +
  ylab("Proportion of Hispanic and Native in GT program")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



In above graph by observing nature of the smooth line we can see that there is under representation of proportion of hispanic and native students in GT program.

```
summarise(schgt,
pr_hn=sum(TOT_HN, na.rm=TRUE) / sum(Tot_enr, na.rm=TRUE),
pr_hn_gt=sum(Tot_HN_GT, na.rm=TRUE) / sum(Tot_stud_gt, na.rm=TRUE))
```

```
##       pr_hn  pr_hn_gt
## 1 0.2808939 0.1908814
```

As we can see above overall proportion of hispanic and native american student in gifted and talented program is less than that of overall proportion of hispanic and native american students. Thus Hispanic and Native American students are under-represented in Gifted & Talented programs.

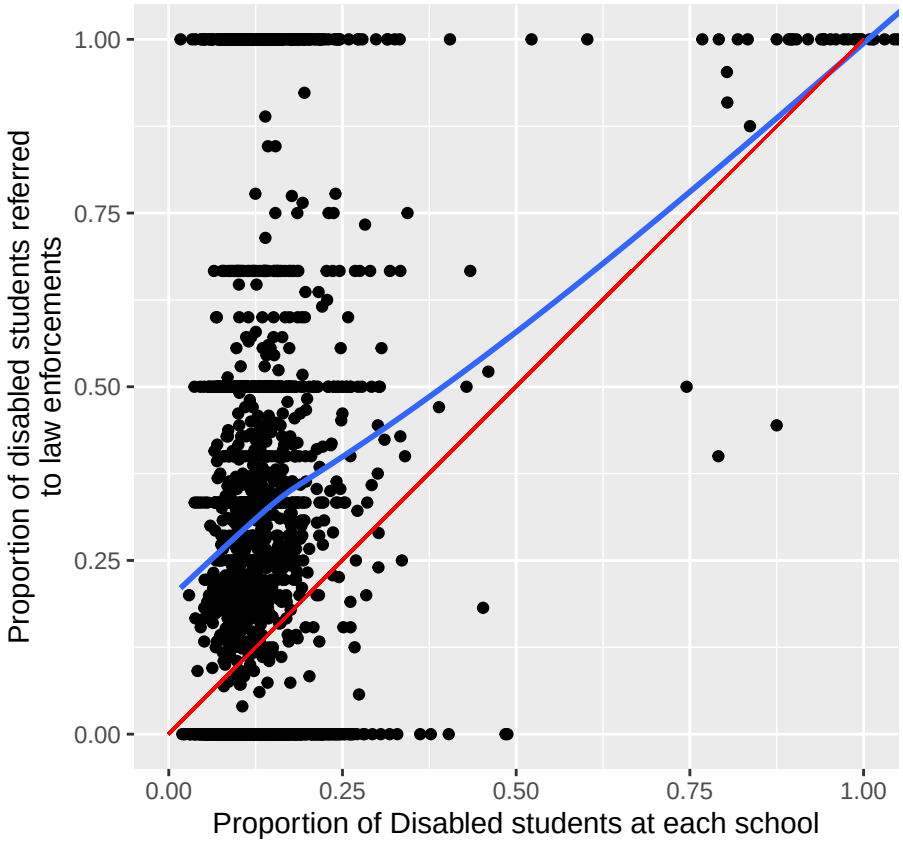
PROBLEM 2

```r
school2<-read.csv("D:/Spring 20 Sem 2/DMP/2015-16-crdc-data (1)/Data Files and Layouts/CRDC 2015-16 Sch
#CREATE A DATAFRAME WITH FOLLOWING

# The total number of students enrolled at each school
# The number of disabled students (served by IDEA) at each school
# The total number of students who were referred to law enforcement
# The number of disabled students (served by IDEA) who were referred to law enforcement
# The proportion of disabled students (served by IDEA) at each school among of all students
# The proportion of students who were referred to law enforcement and are disabled (served by IDEA)amon

sch_disable <- transmute(school2,
enr_tot = TOT_ENR_M + TOT_ENR_F,
enr_dis_tot = TOT_IDEAENR_M + TOT_IDEAENR_F,
tot_law = TOT_DISCWODIS_REF_M + TOT_DISCWODIS_REF_F + TOT_DISCWDIS_REF_IDEA_F + TOT_DISCWDIS_REF_IDEA_M
tot_law_idea = TOT_DISCWDIS_REF_IDEA_F + TOT_DISCWDIS_REF_IDEA_M,
pr_disabled = enr_dis_tot/ enr_tot,
pr_law = tot_law_idea / tot_law)

#PLOT
sch_disable%>%
sample_n(10000) %>%
ggplot(aes(x=pr_disabled, y=pr_law)) +
geom_point() +
geom_smooth(se=FALSE) +
geom_segment(aes(x=0, y=0, xend=1, yend=1), color="red") +
coord_fixed(x=c(0,1), y=c(0,1)) +
labs(x='Proportion of Disabled students at each school', y='Proportion of disabled students referred
     to law enforcements')
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 7877 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 7877 rows containing missing values (geom_point).
```



From the smooth line we can see over representation of proportion of disabled students who were referred to law enforcement. A lot of schools have below 25% percentage of disabled students. Similarly students referred to law enforcements hav maximum population density below 50%.

From the proportions below disabled students are tend to be over represented among disabled students referred to law enforcement.

```
#Calculating overall proportions
summarise(sch_disable,
ovr_pr_dis=sum(enr_dis_tot, na.rm=TRUE) / sum(enr_tot, na.rm=TRUE),
ovr_pr_law=sum(tot_law_idea, na.rm=TRUE) / sum(tot_law, na.rm=TRUE))
```

```
##   ovr_pr_dis ovr_pr_law
## 1  0.1193434  0.2859222
```

PART B

```
library(dbplyr)
```

```
##
## Attaching package: 'dbplyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     ident, sql
```

```
library(DBI)
library(RMySQL)
con<-dbConnect(MySQL(),user="root",password="password",dbname="dblp",host="localhost")
dblp_main<-tbl(con,"general")
dbListTables(con)
```

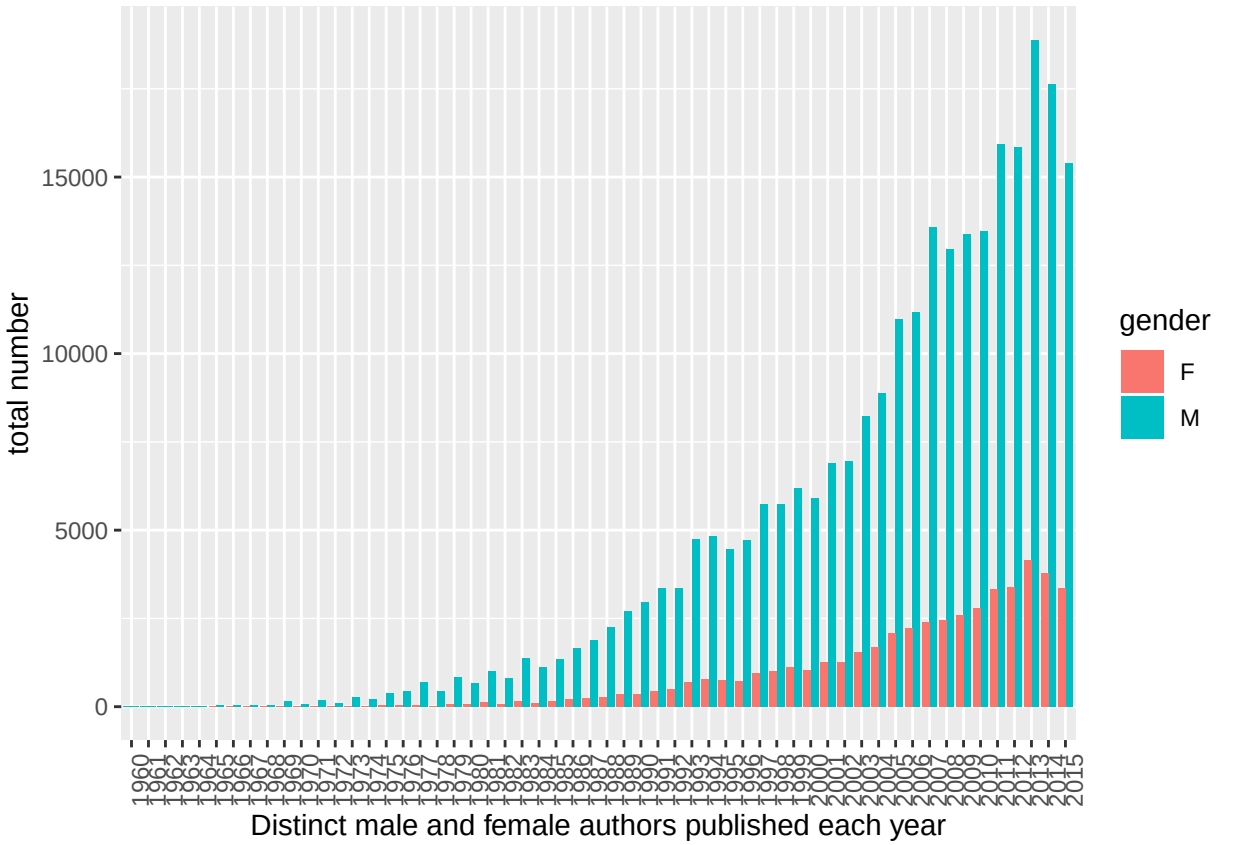
```
## [1] "authors" "general"
```

```
dblp_authors<-tbl(con,"authors")
dblp_authors<-as.data.frame(dblp_authors)
dblp_main<-as.data.frame(dblp_main)
```

PROBLEM 3

From the graph below there is huge difference between male and female authors getting published each year. Male authors are substantially getting increased publications each year. Female authors published are less than 5000 over all the years.

```
dblp_authors_filt<-dblp_authors%>%filter((gender=="M"|gender=="F") & prob>=0.90)
dblp_authors_filt %>%
  inner_join(dblp_main,by=c("k"="k"))%>%
  distinct()%>%
  ggplot()+geom_bar(aes(x=factor(year),fill=gender),position="dodge")+
  xlab("Distinct male and female authors published each year")+
  ylab("total number") + theme(axis.text.x = element_text(angle = 90))
```



PROBLEM 4 From the graph below we can see that proportions of female authors are steadily increasing. For the first few years there were no female candidates. Male candidates are in huge amount compared to female authors.

```
dblp_authors_filt%>%
  inner_join(dblp_main,by=c("k"="k"))%>%
  distinct()%>%
  ggplot()+geom_bar(aes(x=factor(year),fill=gender),position="fill")+
  xlab("Distinct male and female authors published each year")+
  ylab("proportion") + theme(axis.text.x = element_text(angle = 90))
```

PROBLEM 5

In the graph below CS has highest representation of female first authors and SE has lowest representation of female first authors. Also we can see steady increase in female first author in cs unlike other types of conferences. SE also has more number of female authors than before. Some years are missing as there are no first female authors in those years for that type.

```
trial_1 <-
dblp_authors_filt %>%
  inner_join(dblp_main,by=c("k"="k"))%>%
  distinct()%>%filter(pos==0)
trial_2 <- trial_1 %>% gather(cs,de,se,th,key="type",value="count")%>%
  filter(count==1)
trial_3 <- trial_2 %>% group_by(year, type, gender) %>% tally()
trial_4 <- summarise(group_by(trial_3, year, type, gender, n))
```

```r
trial_5_sum_n <- summarise(group_by(trial_4, year), sum(n))
trial_6 <- merge(trial_4, trial_5_sum_n, by = "year")
trial_6 <- trial_6 %>% filter(gender == "F")
trial_7 <- summarise(group_by(trial_6, year, type, gender, n), proportion = n / `sum(n)`)
ggplot(trial_7)+geom_bar(aes(x=factor(year), y=proportion), stat = "identity")+
  facet_wrap(~type) +
  xlab("Years")+
  ylab("Proportion of female with first authorship")+
  theme(axis.text.x = element_text(angle = 90))
```