Comp Sci Final Project

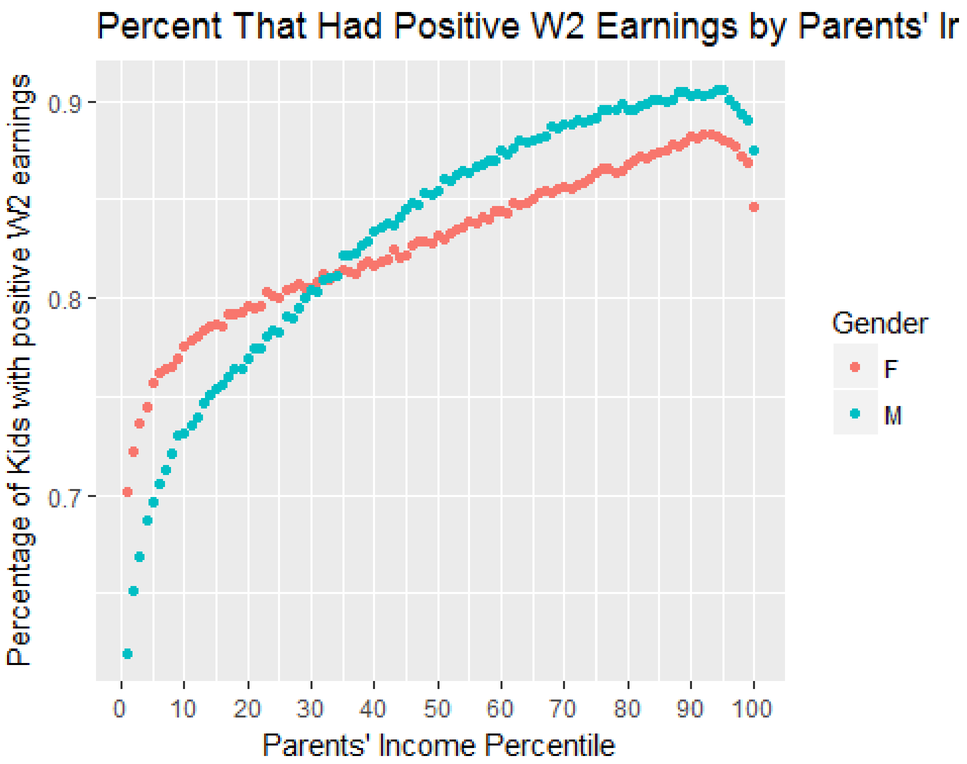
Daisy Richmond, Cory Kieras, Alexandra Rosati, Emily Olson

5/11/2017

```
library(ggplot2)
library(maps)
gender_nat <- read.csv("C:/Users/Owner/Desktop/CSCI 125/Final
Project/gender_nat.csv")
gender_cty <- read.csv("C:/Users/Owner/Desktop/CSCI 125/Final
Project/gender_cty.csv")
```

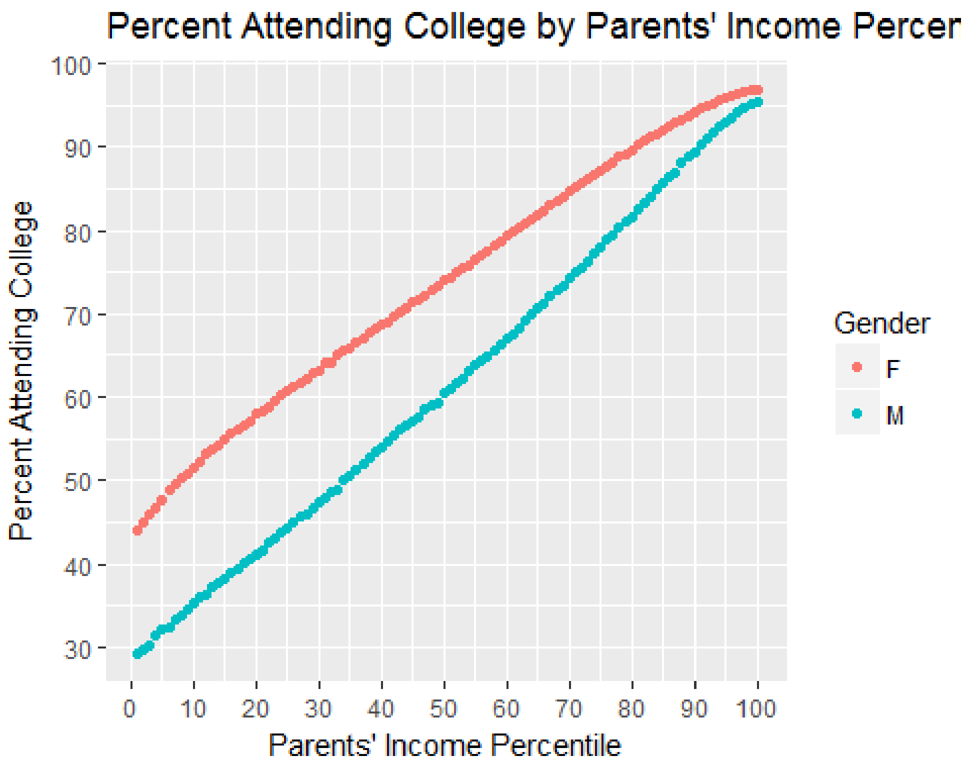
1. First we looked at a graph of the percentage of kids who had positive earning based on their W2 against their parents' income percentile, separated by gender. We thought this would be an important thing to look at because children's income is heavily influenced by their parents' income. We wanted to determine if this effect was any different for males and females. Our graph shows that among parents with low incomes, females were more likely than males to have positive earnings, but among parents with higher incomes, about the 30th percentile and above, males were more likely than females to have positive earnings.

```
gender<-c(rep("F",100),rep("M",100))
percentile<-c(gender_nat$par_pctile,gender_nat$par_pctile)
college<-c(gender_nat$coll1823_f*100,gender_nat$coll1823_m*100)
w2<-c(gender_nat$w2_or_sc_nz_30_f, gender_nat$w2_or_sc_nz_30_m)
kids_individual_rank<-c(gender_nat$kid_indv_rank_30_f,
gender_nat$kid_indv_rank_30_m)
m2<-data.frame(cbind(percentile,college,gender, w2, kids_individual_rank),
stringsAsFactors=F)
#head(m2)
ggplot(data=m2, aes(x=as.numeric(percentile), y=as.numeric(w2),
col=as.factor(gender)))+geom_point()+scale_x_continuous(breaks=seq(0,100,10))
+scale_y_continuous(breaks=seq(0,10,.1))+labs(x="Parents' Income
Percentile",y="Percentage of Kids with positive W2
earnings",color="Gender")+ggtitle("Percent That Had Positive W2 Earnings by
Parents' Income Percentile and Gender of Kid")
```



2. We wanted to investigate the relationship between parents' income and the percent of children that go to college, and to see whether there is a difference for men and women.

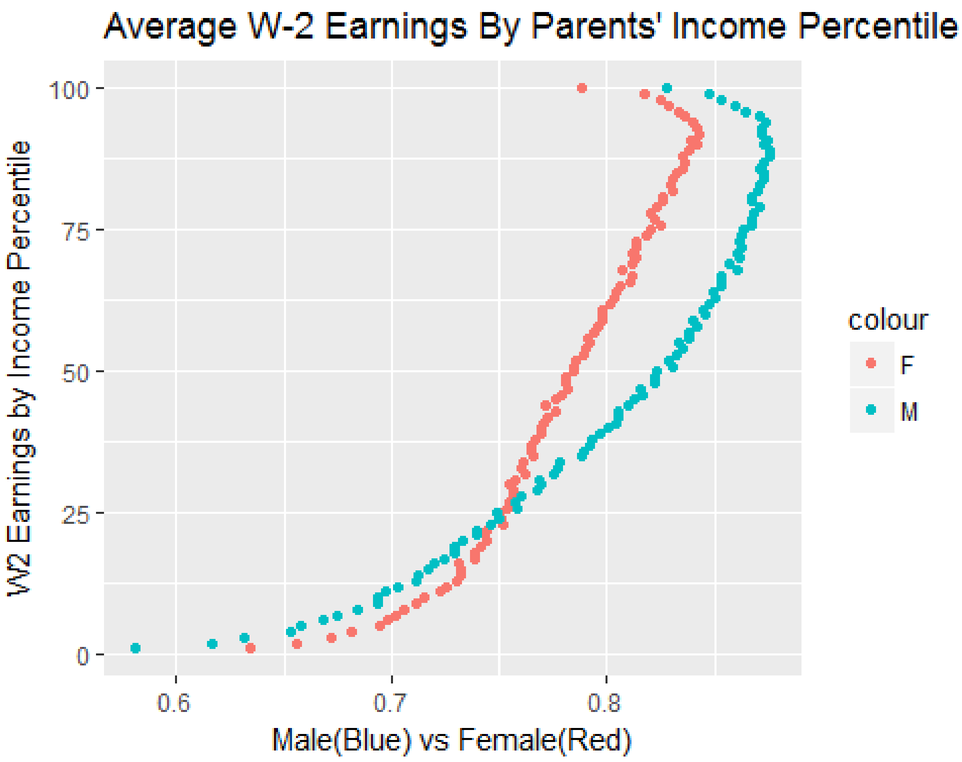
```
college<-c(gender_nat$coll1823_f*100,gender_nat$coll1823_m*100)
gender<-c(rep("F",100),rep("M",100))
percentile<-c(gender_nat$par_pctile,gender_nat$par_pctile)
m2<-data.frame(cbind(percentile,college,gender),stringsAsFactors = F)
#Plot
ggplot(data =
m2,aes(x=as.numeric(percentile),y=as.numeric(college)))+geom_point(aes(color=
factor(gender)))+
  scale_x_continuous(breaks=seq(0,100,10))+
  scale_y_continuous(breaks=seq(0,100,10))+
  labs(x="Parents' Income Percentile",y="Percent Attending College")+
  labs(color="Gender")+
  ggtitle("Percent Attending College by Parents' Income Percentile and
Gender")
```



This shows that more women than men attend college, regardless of parents' income level. However, the gap is largest for low-income people and shrinks considerably at the highest income percentiles.

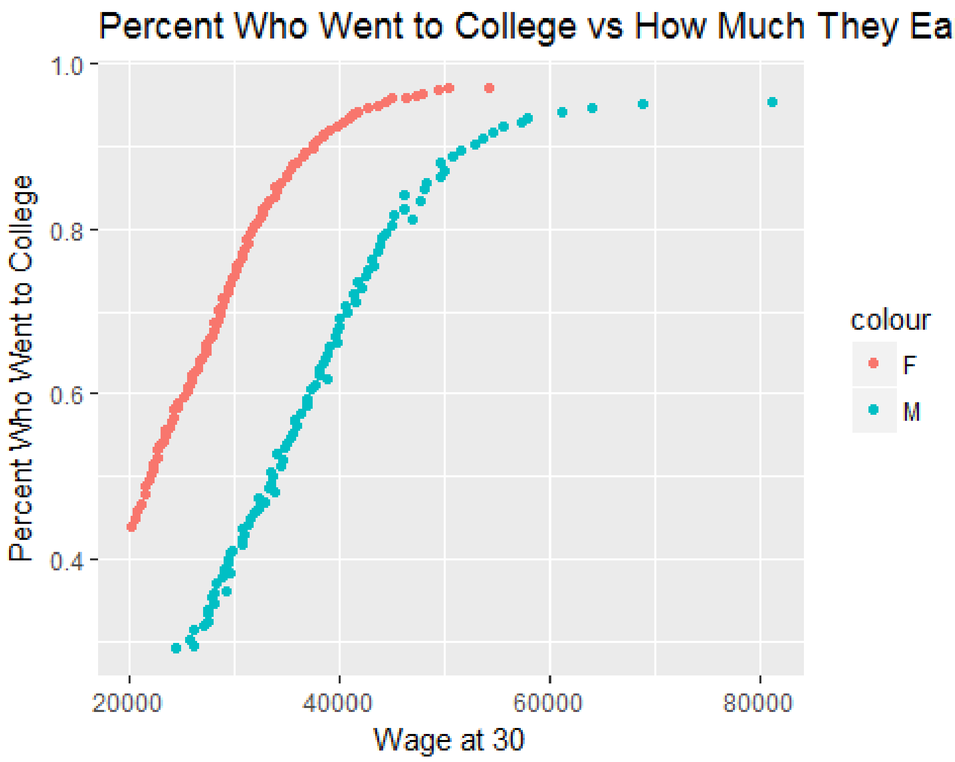
3. We wanted to investigate whether there would be a difference in the income earnings of a parent if they had a male or female. Interestingly, there is a small correlation based on the graph. Parents with a male child tend to make more money than parents with female children.

```
q3<-ggplot()+geom_point(data=gender_nat,
aes(x=w2_pos_30_f,y=par_pctile,color="F"))+geom_point(data=gender_nat,aes(x=w
2_pos_30_m,y=par_pctile,color="M"))+xlab("Male(Blue) vs
Female(Red)")+ylab('W2 Earnings by Income Percentile')+ggtitle("Average W-2
Earnings By Parents' Income Percentile: Male vs Female")
print(q3)
```



4. For this graph we wanted to investigate whether there was a correlation between going to college and W2 earnings when the students are 30. There is, but even if both females and males attend college, males make more. Not only that, but some males who were in the smaller percentile of attending college (so, some didn't) earned more than their female counterparts who overall had a higher rate of college attendance.

```
q4<-ggplot()+geom_point(data=gender_nat,
aes(x=w2wages_30_f,y=coll1823_f,color="F"))+geom_point(data=gender_nat,aes(x=
w2wages_30_m,y=coll1823_m,color="M"))+xlab("Wage at 30")+ylab("Percent Who
Went to College")+ggtitle("Percent Who Went to College vs How Much They Earn
at 30")
print(q4)
```



5. In this question, we wanted to see what the trend was for the differences in US Income Rankings between ages 26-30. We wanted to compare these differences for males and females. To do this, first we determined the differences between males at age 26 and subtracted males at age 30, and did the same with females. We plotted this against the parent income percentile.

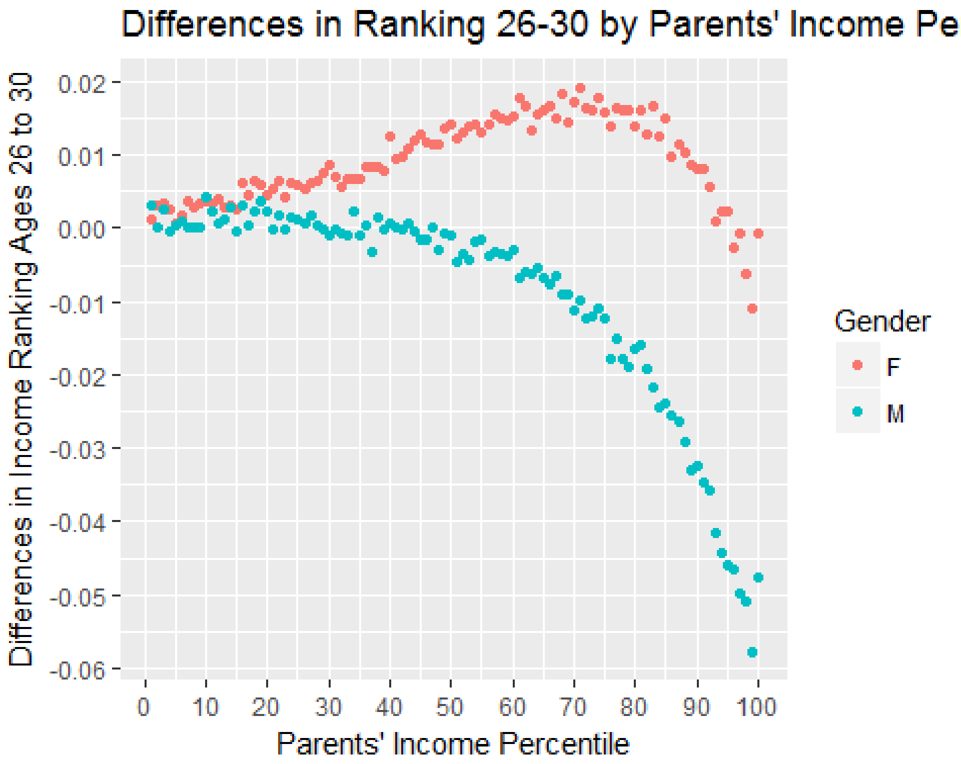
```
femdif<-gender_nat$kid_indv_rank_26_f-gender_nat$kid_indv_rank_30_f
maledif<-gender_nat$kid_indv_rank_26_m-gender_nat$kid_indv_rank_30_m
differences<-data.frame("gender"=gender,"differences"=c(femdif,maledif))
m3<-data.frame(cbind(percentile,college,gender, w2, differences),
stringAsFactors=F)
mean(femdif)

## [1] 0.009131601

mean(maledif)

## [1] -0.009277805

ggplot(data =
m3,aes(x=as.numeric(percentile),y=as.numeric(differences))+geom_point(aes(co
lor=factor(gender)))+
scale_x_continuous(breaks=seq(0,100,10))+
scale_y_continuous(breaks=seq(-1,1,.01))+
labs(x="Parents' Income Percentile",y="Differences in Income Ranking Ages 26
to 30")+
labs(color="Gender")+
ggtitle("Differences in Ranking 26-30 by Parents' Income Percentile and
Gender")
```



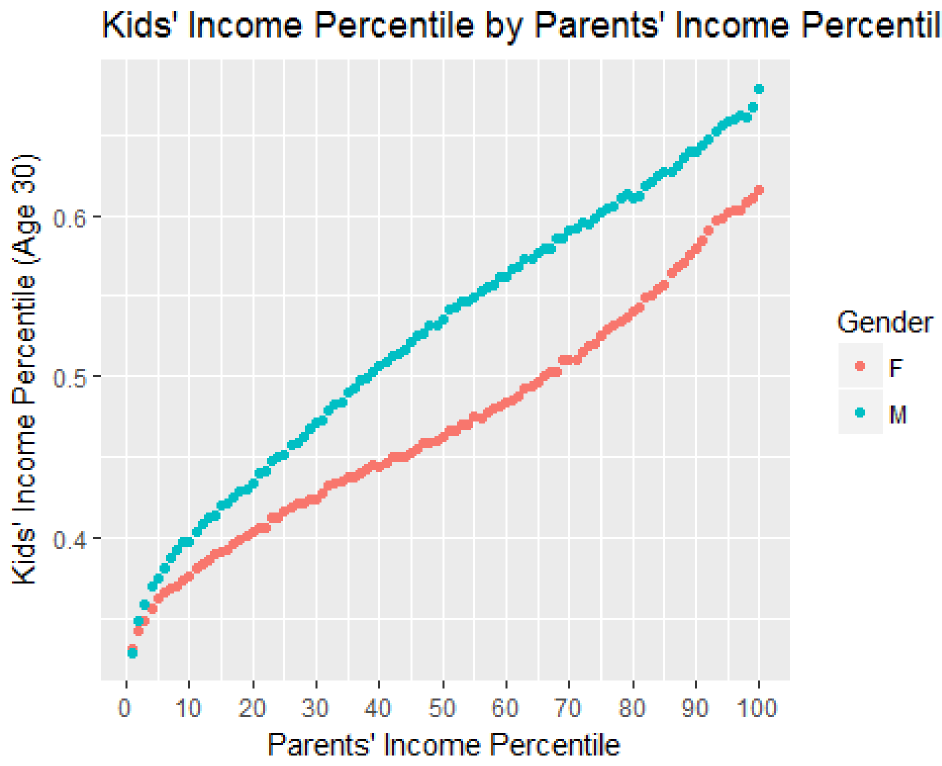
This shows a trend of males having a negative value for the difference in income ranking between 26 and 30 years of age, and women having a positive difference. This means that men tend to have a higher income ranking at age 30 than when they were 26, and women tend to have a lower ranking. However, the differences are very small, and it is possible that these are inconsequential.

6. Similar to our first question, we also looked at kids' income percentile versus parents' income percentile, separated by the gender of the kid. This graph showed us that at every level of parental income percentile, males had a higher income percentile than females.

```
gender<-c(rep("F",100),rep("M",100))
percentile<-c(gender_nat$par_pctile,gender_nat$par_pctile)
w2<-c(gender_nat$w2_or_sc_nz_30_f, gender_nat$w2_or_sc_nz_30_m)
kids_individual_rank<-c(gender_nat$kid_indv_rank_30_f,
gender_nat$kid_indv_rank_30_m)
m2<-data.frame(cbind(percentile,college,gender, w2, kids_individual_rank),
stringsAsFactors = FALSE)
head(m2)
```

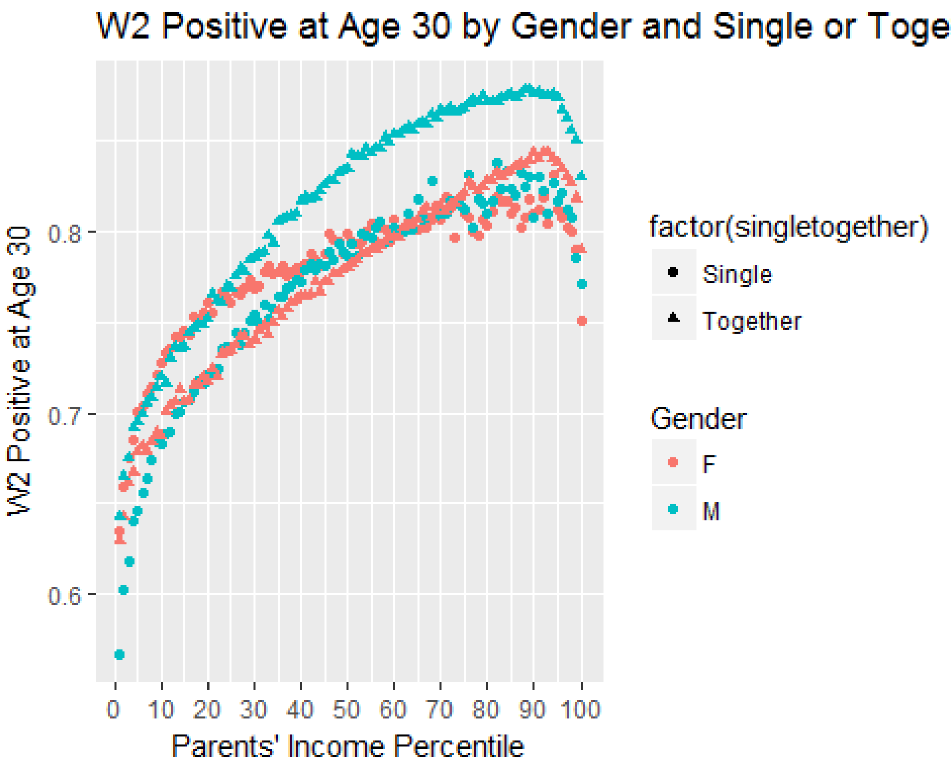
```
##   percentile    college gender          w2 kids_individual_rank
## 1          1  43.887809      F 0.701873005          0.331520241
## 2          2 44.9288888      F 0.722179168            0.3425815
## 3          3 45.9616169      F 0.736014823          0.349440995
## 4          4 46.7593153      F 0.744417547          0.356086894
## 5          5 47.7521773      F   0.7567874          0.363433717
## 6          6 48.7416983      F 0.762119551          0.366766942
```

```
ggplot(data = m2, aes(x=as.numeric(percentile),
y=as.numeric(kids_individual_rank),
col=gender))+geom_point()+scale_x_continuous(breaks=seq(0,100,10))+scale_y_co
ntinuous(breaks=seq(0,10,.1))+labs(x="Parents' Income Percentile",y="Kids'
Income Percentile (Age 30)",color="Gender")+ggtitle("Kids' Income Percentile
by Parents' Income Percentile and Gender of Kid")
```



7. In this question, we wanted to determine if there is any significant disparity in not only the genders' report of their w2s, but also how their parents' marriage status might affect that. In order to do this, we made factors of both gender and marital status, and superimposed them onto one graph in order to see the different trends.

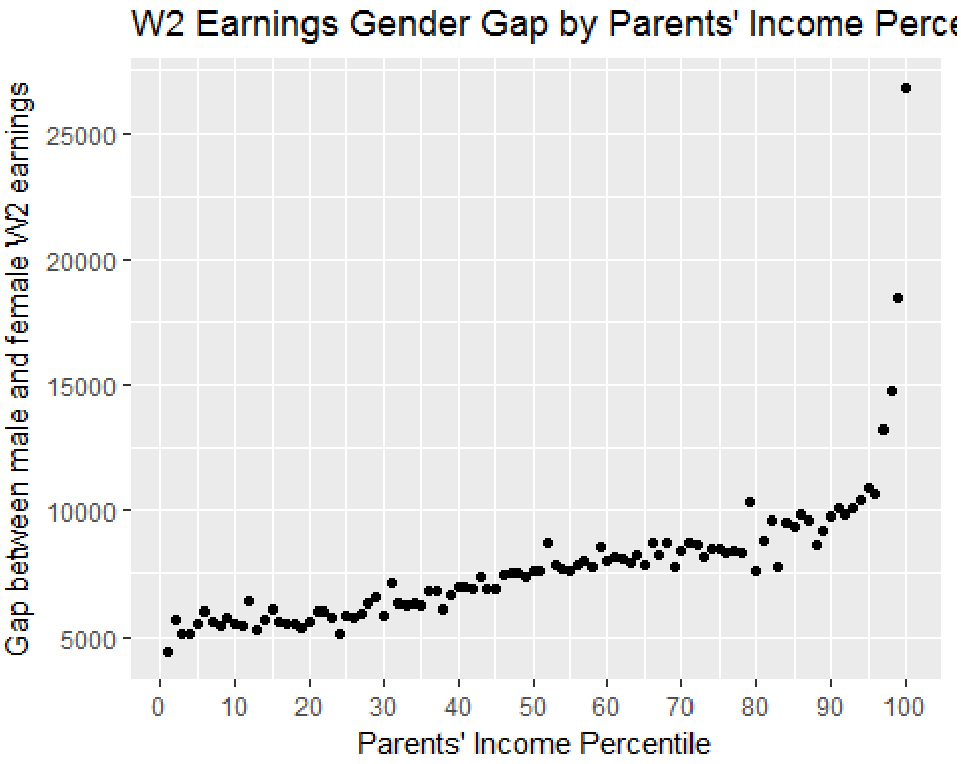
```
singleparents<-c(gender_nat$w2_pos_30_f_sp,gender_nat$w2_pos_30_m_sp)
togetherparents<-c(gender_nat$w2_pos_30_f_tp,gender_nat$w2_pos_30_m_tp)
parents<-c(singleparents,togetherparents)
singletogether<-c(rep("Single",200),rep("Together",200))
gender1<-c(gender,gender)
percentile1<-c(percentile,percentile)
m4<-
data.frame(cbind(percentile1,gender1,parents,singletogether),stringsAsFactors
= FALSE)
ggplot(data =
m4,aes(x=as.numeric(percentile1),y=as.numeric(parents)))+geom_point(aes(color
=factor(gender1),shape=factor(singletogether)))+
scale_x_continuous(breaks=seq(0,100,10))+
scale_y_continuous(breaks=seq(0,10,.10))+
labs(x="Parents' Income Percentile",y="W2 Positive at Age 30")+
labs(color="Gender")+
ggtitle("W2 Positive at Age 30 by Gender and Single or Together Parents")
```



It appears that men who had parents who were together have the highest percentage of positive w2s at age 30. There is very little difference in the trends of the data for male and single parents, female and together parents, and female and single parents. However, it is interesting that the females who had single parents and were below the 50th percentile had a higher percentage of positive w2 when compared to males with single parents and females with together parents.

8. We next wanted to investigate the gap in male and female earnings, and to see whether this gap was affected by how wealthy the people's parents were. We subtracted female earnings from male earnings and plotted this against parents in'come percentile.

```
w2dif<-gender_nat$w2wages_30_m-gender_nat$w2wages_30_f
m8<-data.frame(cbind(gender_nat$par_pctile,w2dif),stringsAsFactors = F)
#Plot
ggplot(data = m8,aes(x=as.numeric(V1),y=as.numeric(w2dif)))+geom_point()+
  scale_x_continuous(breaks = seq(0,100,10))+
  scale_y_continuous(breaks=seq(0,27000,5000))+
  labs(x="Parents' Income Percentile",y="Gap between male and female W2
earnings")+
  ggtitle("W2 Earnings Gender Gap by Parents' Income Percentile")
```

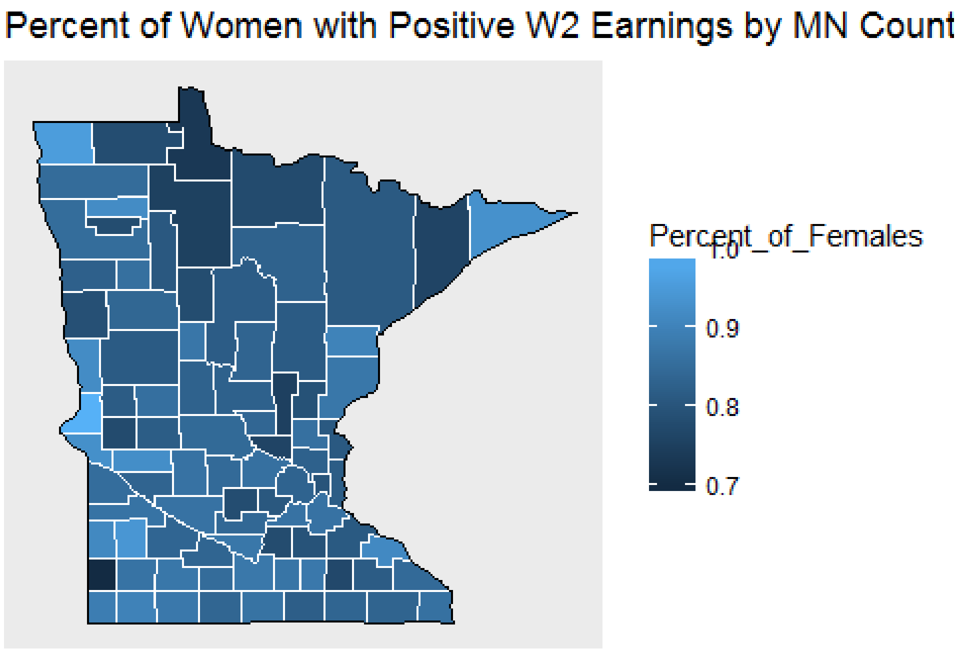


We found a positive wage gap at all levels, indicating that men earned more than women regardless of parent income. This gap increased at higher parental incomes, with large jumps in the gender wage gaps at the very highest incomes.

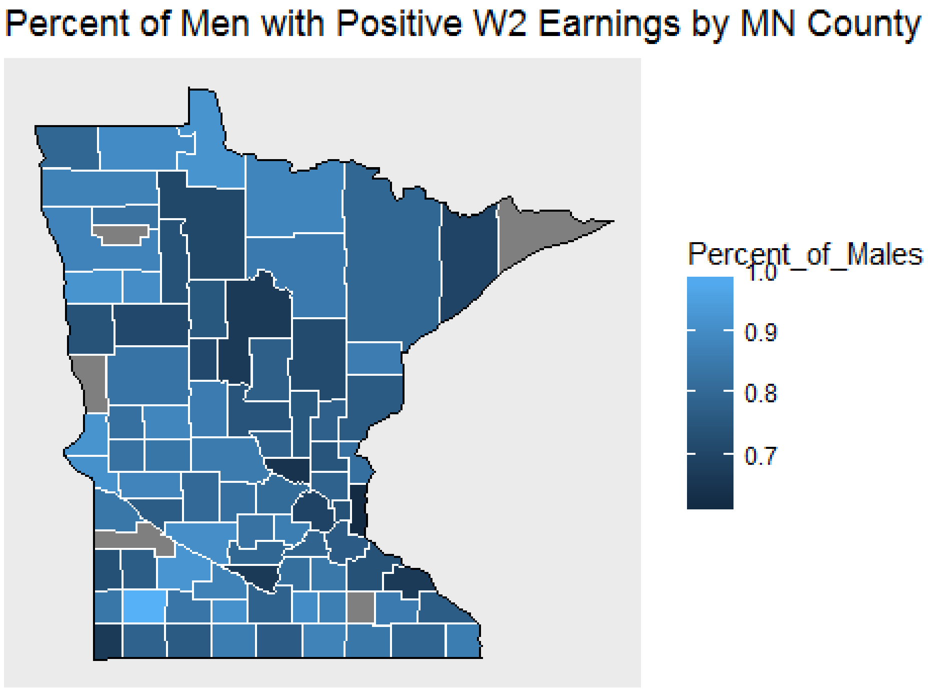
9. Thank you to http://eriqande.github.io/rep-res-web/lectures/making-maps-with-R.html for the very helpful tutorial on making maps

We wanted to know which counties in Minnesota had the highest and lowest percentages of for positive W2 earnings for men and for women. To investigate this we created maps of the percent of women and the percent of men in each county in Minnesota that had positive W2 earnings at age 30.

```
#Subset data since we just want MN
gender_mn<-subset(gender_cty,statename=="Minnesota")
#In order to merge we need county names to be all lower case. Using sapply to
fix entire vector
gender_mn$county_name<-sapply(gender_mn$county_name,tolower)
#Loading map data and subsetting to MN
states<-map_data("state")
mn_df<-subset(states,region=="minnesota")
counties<-map_data("county")
mn_counties<-subset(counties,region=="minnesota")
#Creating base plot of Minnesota, will add additional layers to this
mn_base<-ggplot(data = mn_df,mapping = aes(x=long,y=lat,group=group))+
  coord_fixed(1.3)+
  geom_polygon(color="black",fill="grey")
#mn_base+geom_polygon(data=mn_counties,fill=NA,color="white")+geom_polygon(co
lor="black",fill=NA)
#Merge county latitude/longitude dataset with our gender/wage dataset
mn_final<-merge(mn_counties,gender_mn,by.x = "subregion",by.y =
"county_name")
colnames(mn_final)[24]<-"Percent_of_Females"
colnames(mn_final)[23]<-"Percent_of_Males"
#Creating a theme to eliminate grid in background
ditch_the_axes <- theme(
  axis.text = element_blank(),
  axis.line = element_blank(),
  axis.ticks = element_blank(),
  panel.border = element_blank(),
  panel.grid = element_blank(),
  axis.title = element_blank())
#Creating the maps where counties are colored by percent with positive W2
earnings.
mn_base+geom_polygon(data=mn_final,aes(fill=Percent_of_Females),color="white"
)+
  geom_polygon(color="black",fill=NA)+labs(color="Percent")+
  ditch_the_axes+ggtitle("Percent of Women with Positive W2 Earnings by MN
County")
```



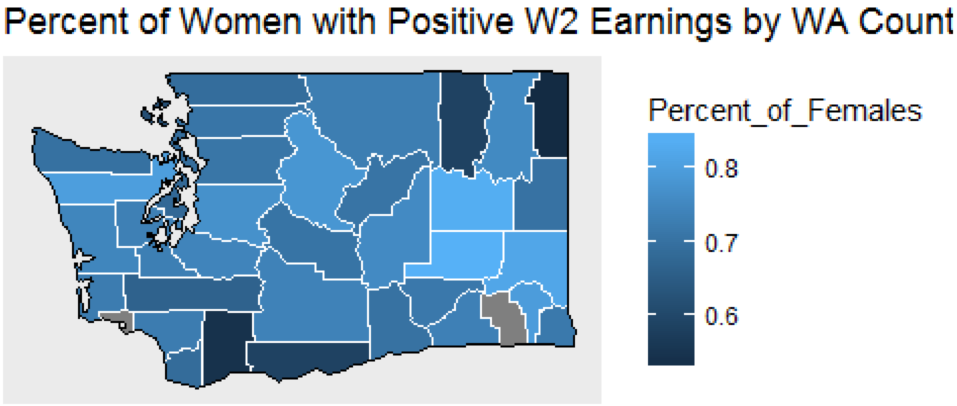
```
mn_base+geom_polygon(data=mn_final,aes(fill=Percent_of_Males),color="white")+
  geom_polygon(color="black",fill=NA)+labs(color="Percent")+
  ditch_the_axes+ggtitle("Percent of Men with Positive W2 Earnings by MN
County")
```



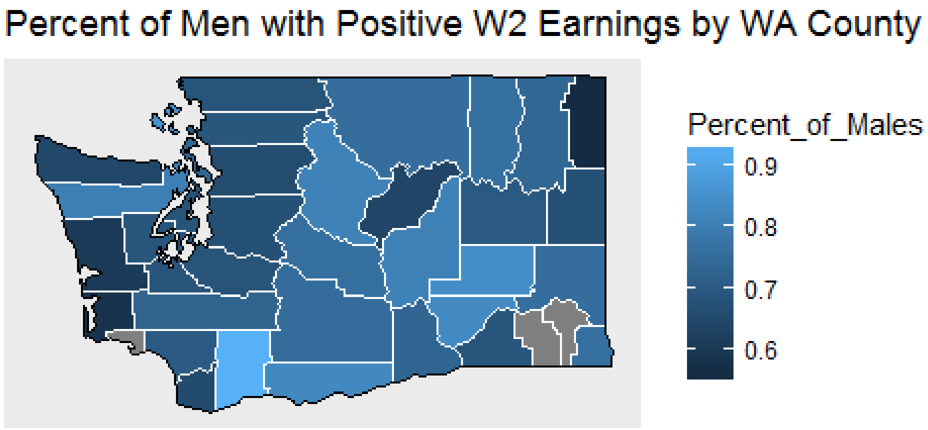
There don't appear to be large differences. Some counties had higher percentage for women and others had higher percentages for men. We also had some missing data for men, so we don't have a complete map.

Cory decided to do the difference between the positive W2 earnings in Washington state because it is where she is from. Here, there isn't a substantial difference either.

```
gender_wa<-subset(gender_cty, statename=="Washington")
gender_wa$county_name<-sapply(gender_wa$county_name,tolower)
states<-map_data("state")
wa_df<-subset(states,region=="washington")
counties<-map_data("county")
wa_counties<-subset(counties,region=="washington")
wa_base<-ggplot(data = wa_df,mapping =
aes(x=long,y=lat,group=group))+coord_fixed(1.3)+geom_polygon(color="black",fi
ll="grey")
#wa_base+geom_polygon(data=wa_counties,fill=NA,color="white")+geom_polygon(co
lor="black",fill=NA)
wa_final<-merge(wa_counties,gender_wa,by.x="subregion",by.y="county_name")
colnames(wa_final)[24]<-"Percent_of_Females"
colnames(wa_final)[23]<-"Percent_of_Males"
ditch_the_axes <- theme(
  axis.text = element_blank(),
  axis.line = element_blank(),
  axis.ticks = element_blank(),
  panel.border = element_blank(),
  panel.grid = element_blank(),
  axis.title = element_blank())
wa_base+geom_polygon(data=wa_final,aes(fill=Percent_of_Females),color="white"
)+
  geom_polygon(color="black",fill=NA)+labs(color="Percent")+
  ditch_the_axes+ggtitle("Percent of Women with Positive W2 Earnings by WA
County")
```

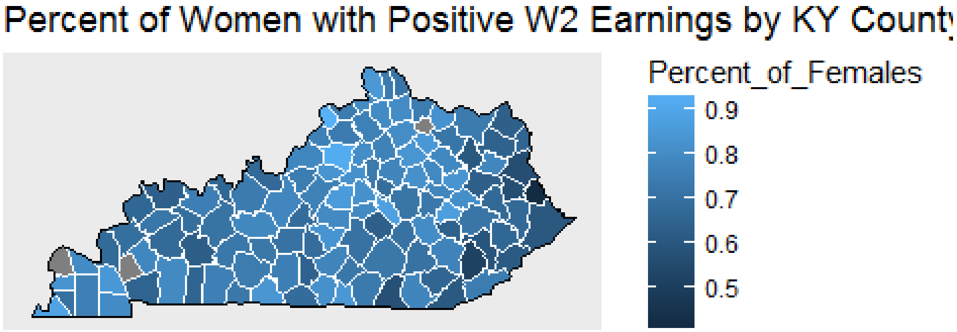


```
wa_base+geom_polygon(data=wa_final,aes(fill=Percent_of_Males),color="white")+
  geom_polygon(color="black",fill=NA)+labs(color="Percent")+
  ditch_the_axes+ggtitle("Percent of Men with Positive W2 Earnings by WA
County")
```

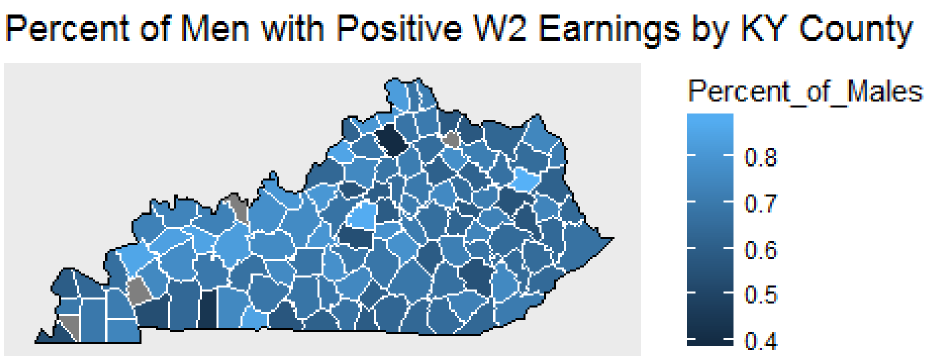


For this question, we wanted to get exemplary status for doing the same mapping technique to the state of Kentucky. We did this because think that this is a good way to compare to another state, and also, because our professor is from Kentucky.

```
gender_ky<-subset(gender_cty,statename=="Kentucky")
gender_ky$county_name<-sapply(gender_ky$county_name,tolower)
states<-map_data("state")
ky_df<-subset(states,region=="kentucky")
counties<-map_data("county")
ky_counties<-subset(counties,region=="kentucky")
ky_base<-ggplot(data = ky_df,mapping = aes(x=long,y=lat,group=group))+
coord_fixed(1.3)+
geom_polygon(color="black",fill="grey")
#ky_base+geom_polygon(data=ky_counties,fill=NA,color="white")+geom_polygon(co
lor="black",fill=NA)
ky_final<-merge(ky_counties,gender_ky,by.x = "subregion",by.y =
"county_name")
colnames(ky_final)[24]<-"Percent_of_Females"
colnames(ky_final)[23]<-"Percent_of_Males"
ditch_the_axes <- theme(
axis.text = element_blank(),
axis.line = element_blank(),
axis.ticks = element_blank(),
panel.border = element_blank(),
panel.grid = element_blank(),
axis.title = element_blank())
ky_base+geom_polygon(data=ky_final,aes(fill=Percent_of_Females),color="white"
)+
geom_polygon(color="black",fill=NA)+labs(color="Percent")+
ditch_the_axes+ggtitle("Percent of Women with Positive W2 Earnings by KY
County")
```



```
ky_base+geom_polygon(data=ky_final,aes(fill=Percent_of_Males),color="white")+
geom_polygon(color="black",fill=NA)+labs(color="Percent")+
ditch_the_axes+ggtitle("Percent of Men with Positive W2 Earnings by KY
County")
```



There doesn't appear to be a substantial difference in instances of positive w2s between men and women in Kentucky. Although some of the counties show women with a lesser percentage than men, others show women with a higher percentage.