Poverty Data - Exploratory Analysis

Michael Bemus

2024-09-26

Here, we want to examine some of the trends occurring in the data.

Load Data and Packages

```
df <- read.csv("SubsidyClass/dataset.csv")
```

We have a few data processing steps that we have to take before we begin. First, our categorical variables have been read in as character data, so we need to retransform them into factors.

```
# Convert all of these variables to factors.
df$sex <- factor(df$sex)
df$mar <- factor(df$mar)
df$hispanic <- factor(df$hispanic)
df$race <-factor(df$race)
df$edu <- factor(df$edu)
df$off_pov <- factor(df$off_pov)
df$spm_pov <- factor(df$spm_pov)
df$ui_kids <- factor(df$ui_kids)
df$mortgage <- factor(df$mortgage)
df$st <- factor(df$st)
df$group <- factor(df$group)
```

Now, we have a few new variables we want to explore. First, we want to condense our poverty classifications into 1 column to simplify visualization processes.

```
# Initialize new column as non-impoverished.
df$pov <- "Neither"

# Classify those who are exclusively impoverished by the federal def.
df$pov[(df$off_pov == 1)&(df$spm_pov == 0)] <- "Official"

# Classify those who are exclusively impoverished by the supplemental def.
df$pov[(df$off_pov == 0)&(df$spm_pov == 1)] <- "Supplemental"

# Classify those who fall under both categories of impoverishment.
df$pov[(df$off_pov == 1)&(df$spm_pov == 1)] <- "Both"

# Transform new column to a factor.
df$pov <- factor(df$pov)

head(df$pov, 10)
```

```
##  [1] Neither  Neither  Neither  Neither  Neither  Neither  Neither  Official
##  [9] Neither  Neither
## Levels: Both Neither Official Supplemental
```

Next, we know how much an individual would have paid before and after their tax credit was applied. We want to change the fed_tax_bc column to contain an individual's tax credit.

```
# Compute credit by subtracting before-credit from after-credit tax.
df$fed_tax_bc <- df$fed_tax_bc - df$fed_tax

# Rename column to reflect the change.
names(df)[names(df) == 'fed_tax_bc'] <- 'tax_credit'

# Check if any values in the new column are negative.
length(df$tax_credit[df$tax_credit < 0])
```

```
## [1] 17216
```

The 17216 observations seen with negative credit values should be examined further to see if they have any errors. Most likely, they could represent a backwards input between the two columns. Forcing these occurrences to be positive is likely appropriate for analysis.

Lastly, since we are interested in the subsidies earned by individuals, we wnat to see how much total subsidy each observation has earned.

```
# Add up all subsidy variables.
df$ttl_sub <- df$snap + df$house_sub + df$slunch + df$energy + df$wic
```

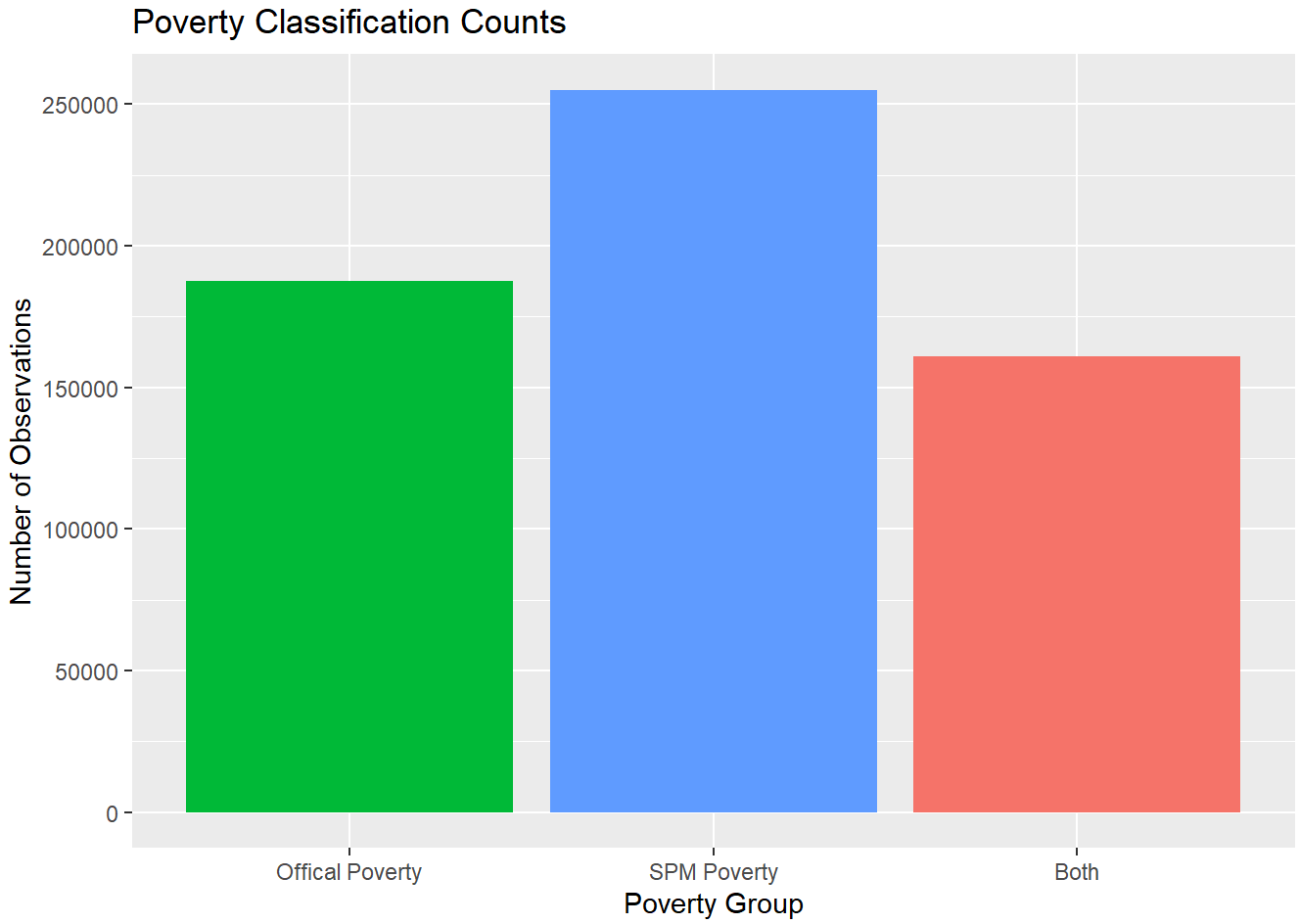
Univariate Visualizations

Now that we've processed the data, we can start to create exploratory visualizations.

Poverty Definition Distributions

We'll start with examining how many individuals we have under each poverty category.

```
# We can't do a normal histogram for this, so we need to play with the code.
bar_labs <- c("Offical Poverty", "SPM Poverty", "Both")

# Take counts of how many observations are impoverished under each class.
bar_vals <- c(length(df$off_pov[df$off_pov == 1]), length(df$spm_pov[df$spm_pov == 1]), length(d
f$group[(df$off_pov == 1)&(df$spm_pov == 1)]))

# Create a bar graph.
ggplot(mapping=aes(x=bar_labs, y=bar_vals, fill=bar_labs)) +
  geom_bar(stat="identity") +
  labs(title="Poverty Classification Counts",
       x ="Poverty Group",
       y = "Number of Observations")  +
  theme(legend.position = "none") +
  scale_x_discrete(limits = bar_labs) +
  scale_y_continuous(breaks=c(0, 50000, 100000, 150000, 200000, 250000, 300000))
```



Here, we see that the SPM group has more observations than the official poverty group. In this visualization, we allow for the “Both” class individuals to contribute to official and supplemental, meaning the both column represents the overlap between the two. It is interesting how few observations seem to be exclusively classified under the official metric relative to the supplemental group. This could suggest the official metric is not as useful as the supplemental metric.

From a quantitative perspective, we can look at the counts of each.

```
# First, print the counts for the bar chart above.
for (i in 1:3) {
  print(paste("Count in ", bar_labs[i], ": ", bar_vals[i], sep=""))
}
```

```
## [1] "Count in Offical Poverty: 187497"
## [1] "Count in SPM Poverty: 255063"
## [1] "Count in Both: 160933"
```

```
# Next, we compare the total number of impoverished individuals compared to
# individuals that are not impoverished.
print(paste("Count Total Poverty:", length(df$group[df$group == "poor"])))
```

```
## [1] "Count Total Poverty: 281627"
```

```
print(paste("Count Not Impoverished:", length(df$pov[df$pov == "Neither"])))
```

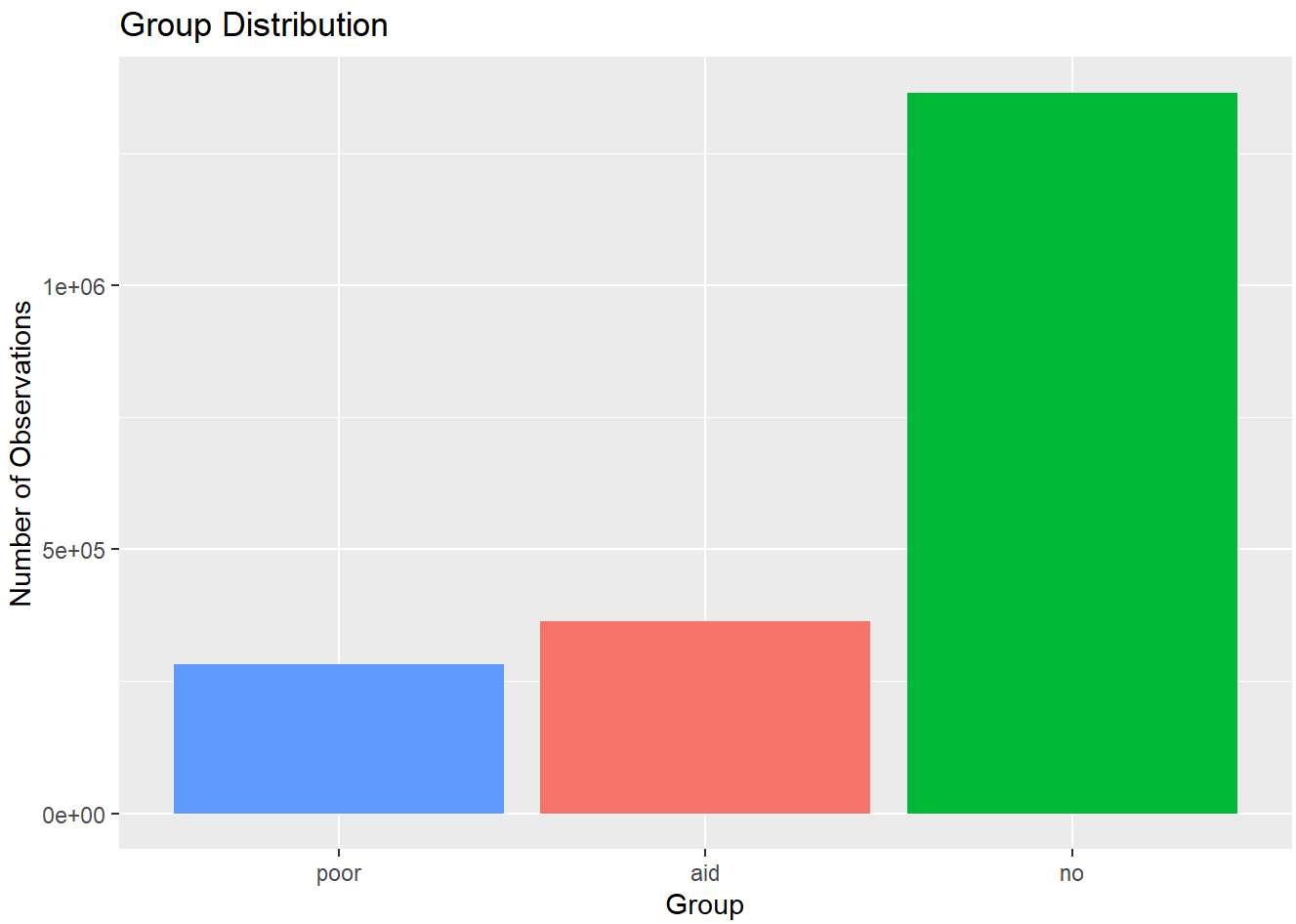
```
## [1] "Count Not Impoverished: 1729348"
```

We see that the SPM metric covers most of our impoverished individuals, with only about 25,000 observations being identified uniquely by the official metric.

Target Group Distributions

Next, we can look at the distribution for our variable of interest.

```
# Create a bar plot comparing the counts of our three classes of interest.
ggplot(df, aes(x=group, fill=group)) + geom_bar() +
  labs(title="Group Distribution", x ="Group", y = "Number of Observations") +
  theme(legend.position = "none") +
  scale_x_discrete(limits = c("poor", "aid", "no"))
```



If we look at our three groups, we see that the “No” group clearly has the most observations, easily doubling the amounts in the other two classes.

```
# Use Group column to calculate and print total percent of each class.
for (i in c("poor", "aid", "no")) {
  print(paste("Percent in group ", i, ": ",
              length(df$group[df$group == i])/length(df$group),
              sep=""))
}
```

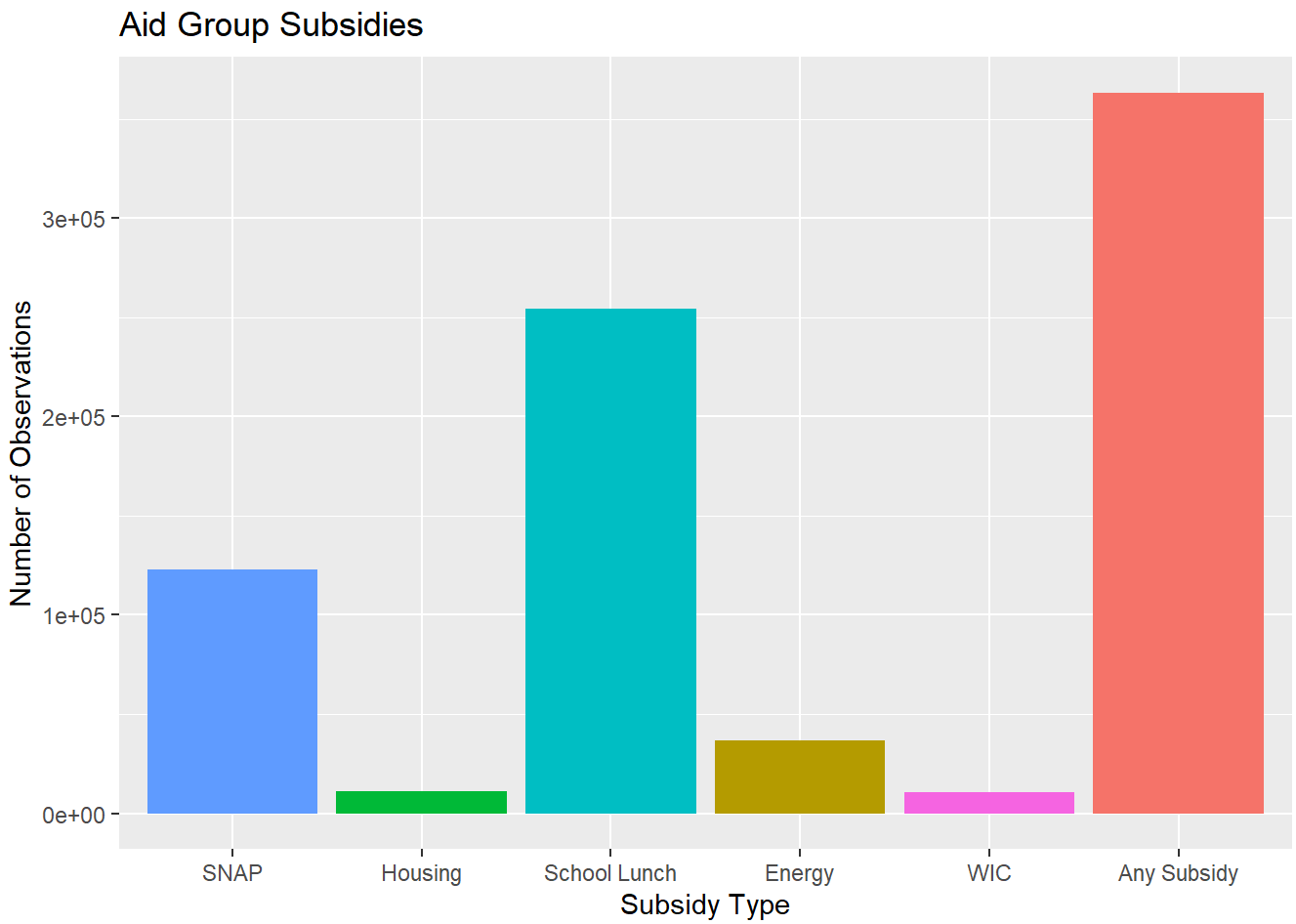
```
## [1] "Percent in group poor: 0.140045003045786"
## [1] "Percent in group aid: 0.180625318564378"
## [1] "Percent in group no: 0.679329678389836"
```

Quantitatively, we see that the Non-Impoverished group contains about 2/3’s of our observations. Of what remains, we have almost 30% more observations in the aid group compared to the poor group.

Subsidy Distributions

The next most interesting variables to examine will be our subsidy variables. We want to compare these across our three groups to see if there is any variance in how subsidies are distributed among them.

```
# Select only those observations in our aid group.
aid_group <- filter(df, group=="aid")

# Find the total number in the aid group who earned subsidies..
aid_w_sub <- length(df$pov[(df$group == "aid")&((df$snap > 0)|
                                                (df$house_sub > 0)|
                                                (df$slunch > 0)|
                                                (df$wic > 0)|
                                                (df$energy > 0))])

# Create the values for our bar chart by counting observations.
bar_vals <- c(length(aid_group$snap[aid_group$snap > 0]),
              length(aid_group$house_sub[aid_group$house_sub > 0]),
              length(aid_group$slunch[aid_group$slunch > 0]),
              length(aid_group$energy[aid_group$energy > 0]),
              length(aid_group$wic[aid_group$wic > 0]), aid_w_sub)

# List out the labels we need for the x-axis.
bar_labs <- c("SNAP", "Housing", "School Lunch", "Energy", "WIC", "Any Subsidy")

# Create our bar plot.
ggplot(mapping=aes(x=bar_labs, y=bar_vals, fill=bar_labs)) +
  geom_bar(stat="identity") +
  labs(title="Aid Group Subsidies",
       x ="Subsidy Type",
       y = "Number of Observations")  +
  theme(legend.position = "none") +
  scale_x_discrete(limits = bar_labs)
```



We see that most subsidies earned by this group are school lunch subsidies. The Energy, Housing, and WIC subsidies are much rarer compared to the other groups. While we do see quite a few earning SNAP subsidies, the value seems to be less than half that of the school lunch subsidies.

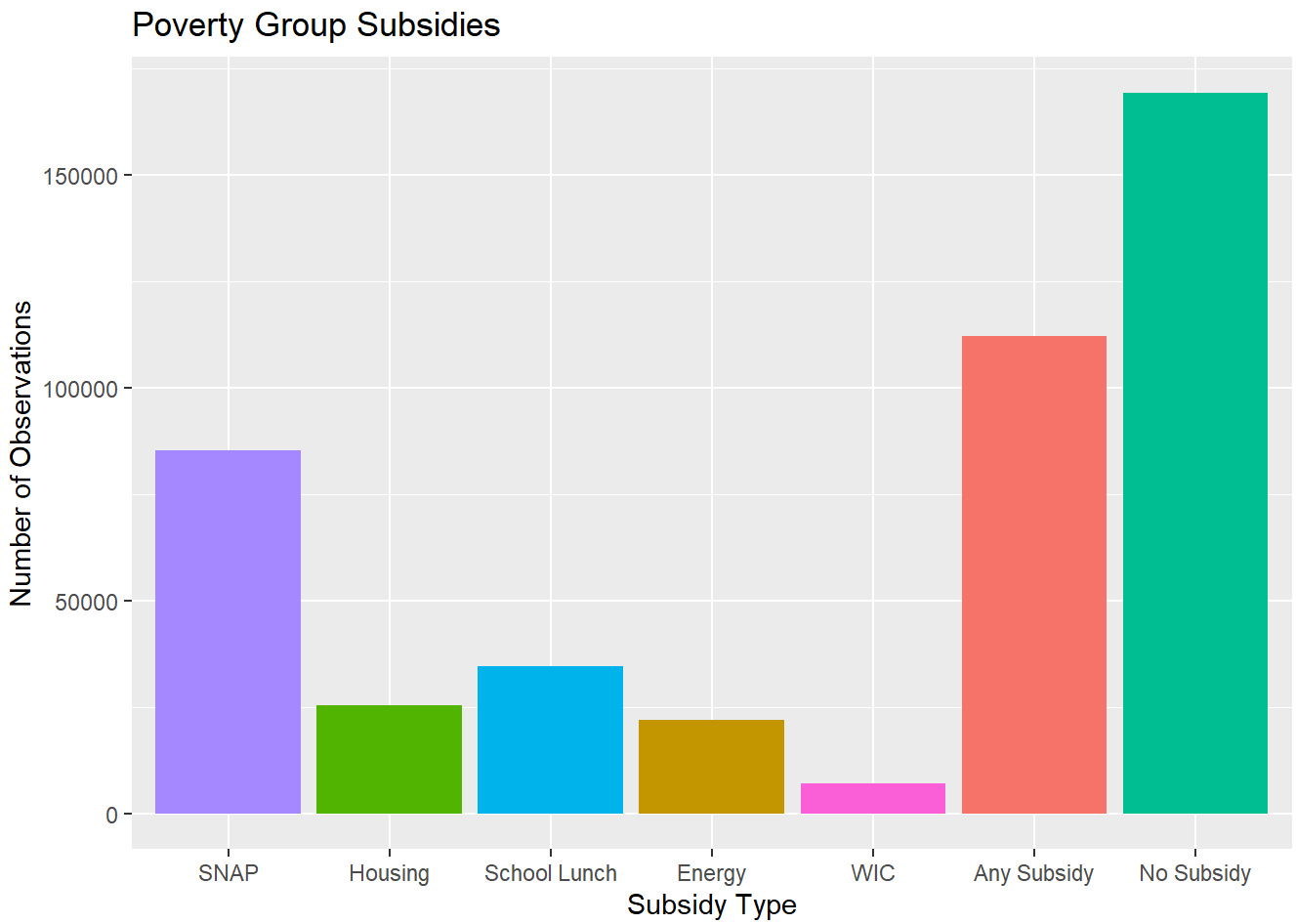
```
# Iterate through last block's calculations to print results.
for (i in 1:5) {
  print(paste("Aid Group Count of ", bar_labs[i], ": ", bar_vals[i], sep=""))
}
```

```
## [1] "Aid Group Count of SNAP: 122649"
## [1] "Aid Group Count of Housing: 11261"
## [1] "Aid Group Count of School Lunch: 254370"
## [1] "Aid Group Count of Energy: 36476"
## [1] "Aid Group Count of WIC: 10538"
```

Seeing the counts, we can verify that school lunch subsidies made up the majority of subsidies earned by this group.

Next, we can copy the above computations for the Poor group.

```
# Select only those observations in our poverty group.
pov_group <- filter(df, group=="poor")

# Count the number of observations not earning subsidies.
pov_no_sub <- length(df$pov[(df$group == "poor")&(df$snap == 0)&
                            (df$house_sub == 0)&(df$slunch == 0)&
                            (df$wic == 0)&(df$energy == 0)])

# Count the number of observations earning subsidies.
pov_w_sub <- length(df$pov[(df$group == "poor")&((df$snap > 0)|
                                                 (df$house_sub > 0)|
                                                 (df$slunch > 0)|
                                                 (df$wic > 0)|
                                                 (df$energy > 0))])

# Create a list of bar values for each subsidy type.
bar_vals <- c(length(pov_group$snap[pov_group$snap > 0]),
              length(pov_group$house_sub[pov_group$house_sub > 0]),
              length(pov_group$slunch[pov_group$slunch > 0]),
              length(pov_group$energy[pov_group$energy > 0]),
              length(pov_group$wic[pov_group$wic > 0]),
              pov_w_sub, pov_no_sub)

# Create our bar chart labels.
bar_labs <- c("SNAP", "Housing", "School Lunch", "Energy", "WIC",
              "Any Subsidy", "No Subsidy")

# Create a bar plot of the above data.
ggplot(mapping=aes(x=bar_labs, y=bar_vals, fill=bar_labs)) +
  geom_bar(stat="identity") +
  labs(title="Poverty Group Subsidies",
       x ="Subsidy Type",
       y = "Number of Observations")  +
  theme(legend.position = "none") +
  scale_x_discrete(limits = bar_labs)
```



Immediately, what stands out to me in this chart is that more individuals who are known to be impoverished are not receiving any of these subsidies than those who are. Participation in this survey requires an individual to have a home address, so these people can be found and contacted by the government. The fact that they are not earning subsidies likely represents some level of failing in our subsidy system.

We see that the subsidy most impoverished individuals are earning are SNAP subsidies, which makes sense. We still do not see that many of the total population earning housing, energy, or WIC subsidies. Could these programs be expanded?

It is also interesting to see that these individuals receive less school lunch subsidies overall. This is likely a statement of the fact that impoverished individuals cannot afford to have children.

```
# Print out the counts of individuals earning each subsidy type.
for (i in 1:5) {
  print(paste("Poverty Group Count of ", bar_labs[i], ": ", bar_vals[i], sep=""))
}
```

```
## [1] "Poverty Group Count of SNAP: 85220"
## [1] "Poverty Group Count of Housing: 25324"
## [1] "Poverty Group Count of School Lunch: 34573"
## [1] "Poverty Group Count of Energy: 21841"
## [1] "Poverty Group Count of WIC: 7107"
```

```
# Print the counts of individuals earning subsidies vs not earning subsidies.
print(paste("Poverty Group w/ Any Susbdiy:", pov_w_sub))
```

```
## [1] "Poverty Group w/ Any Susbdiy: 112233"
```

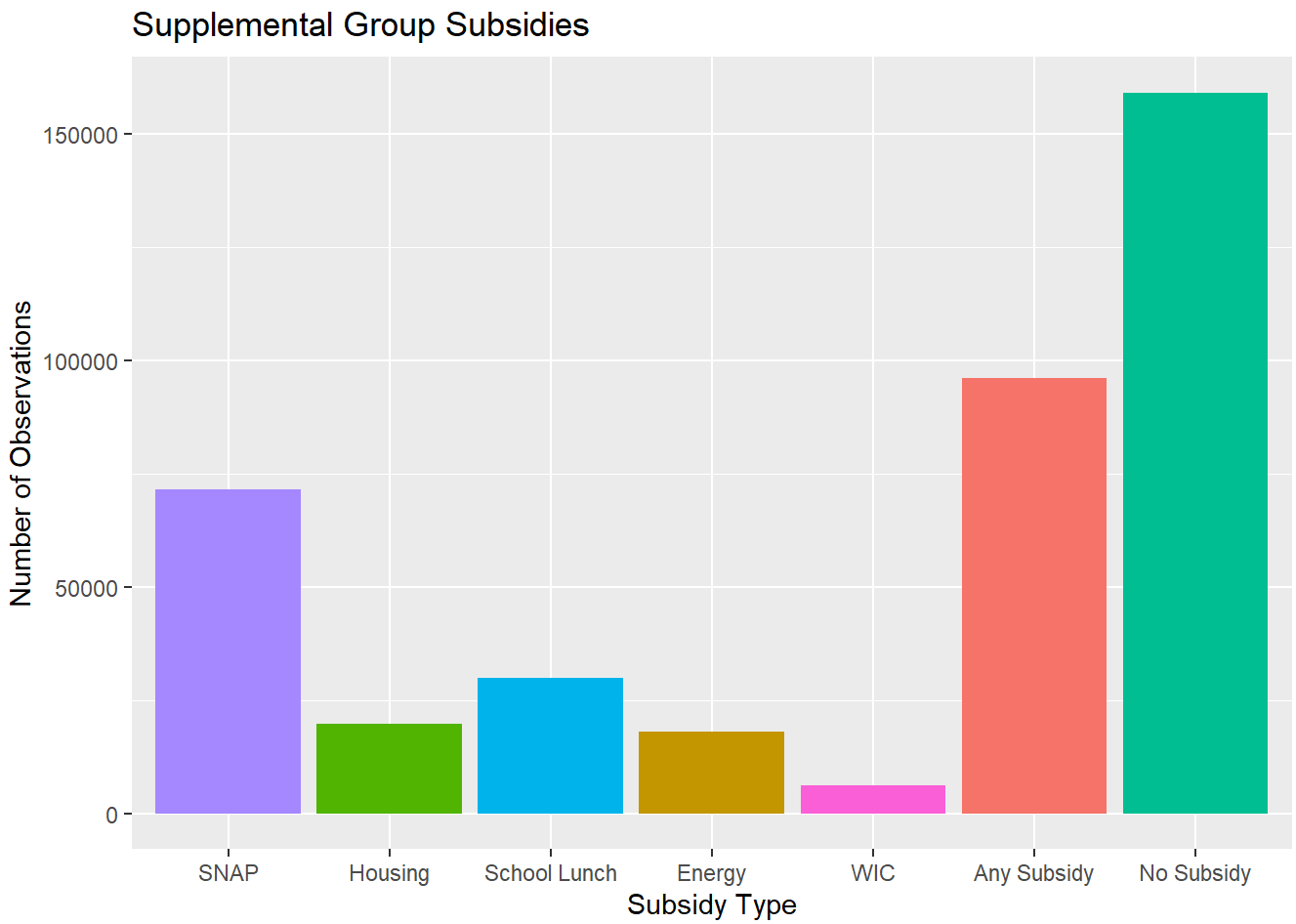
```
print(paste("Poverty Group Count of No Subsidies:", pov_no_sub))
```

```
## [1] "Poverty Group Count of No Subsidies: 169394"
```

Here, we confirm much of what we saw in the above bar chart. Surprisingly, almost 33% more impoverished individuals are not earning subsidies compared to those who are.

We can break this analysis down further into our two poverty groups. We begin with the larger supplemental poverty group.

```
# Select only observations in the supplemental poverty group.
spm_group <- filter(df, spm_pov==1)

# Count how many are not earning subsidies.
spm_no_sub <- length(df$pov[(df$spm_pov==1)&(df$snap == 0)&(df$house_sub == 0)&
                              (df$slunch == 0)&(df$wic == 0)&(df$energy == 0)])

# Count how many are earning subsidies.
spm_w_sub <- length(df$pov[(df$spm_pov==1)&((df$snap > 0)|
                                            (df$house_sub > 0)|
                                            (df$slunch > 0)|
                                            (df$wic > 0)|
                                            (df$energy > 0))])

# Assign bar chart values by counting the number earning each subsidy type.
bar_vals <- c(length(spm_group$snap[spm_group$snap > 0]),
              length(spm_group$house_sub[spm_group$house_sub > 0]),
              length(spm_group$slunch[spm_group$slunch > 0]),
              length(spm_group$energy[spm_group$energy > 0]),
              length(spm_group$wic[spm_group$wic > 0]),
              spm_w_sub, spm_no_sub)

# Create our bar chart labels.
bar_labs <- c("SNAP", "Housing", "School Lunch", "Energy", "WIC",
              "Any Subsidy", "No Subsidy")

# Plot the above data.
ggplot(mapping=aes(x=bar_labs, y=bar_vals, fill=bar_labs)) +
  geom_bar(stat="identity") +
  labs(title="Supplemental Group Subsidies",
       x ="Subsidy Type",
       y = "Number of Observations")  +
  theme(legend.position = "none") +
  scale_x_discrete(limits = bar_labs)
```



We see that this distribution follows a very similar pattern to our overall poverty class.

```
# Iterate through bar chart values to print counts.
for (i in 1:5) {
  print(paste("Supplemental Group Count of ", bar_labs[i], ": ", bar_vals[i],
              sep=""))
}
```

```
## [1] "Supplemental Group Count of SNAP: 71437"
## [1] "Supplemental Group Count of Housing: 19632"
## [1] "Supplemental Group Count of School Lunch: 29879"
## [1] "Supplemental Group Count of Energy: 17905"
## [1] "Supplemental Group Count of WIC: 6051"
```

```
# Compare the count earning subsidies compared to those who aren't.
print(paste("Supplemental Group w/ Any Susbdiy:", spm_w_sub))
```

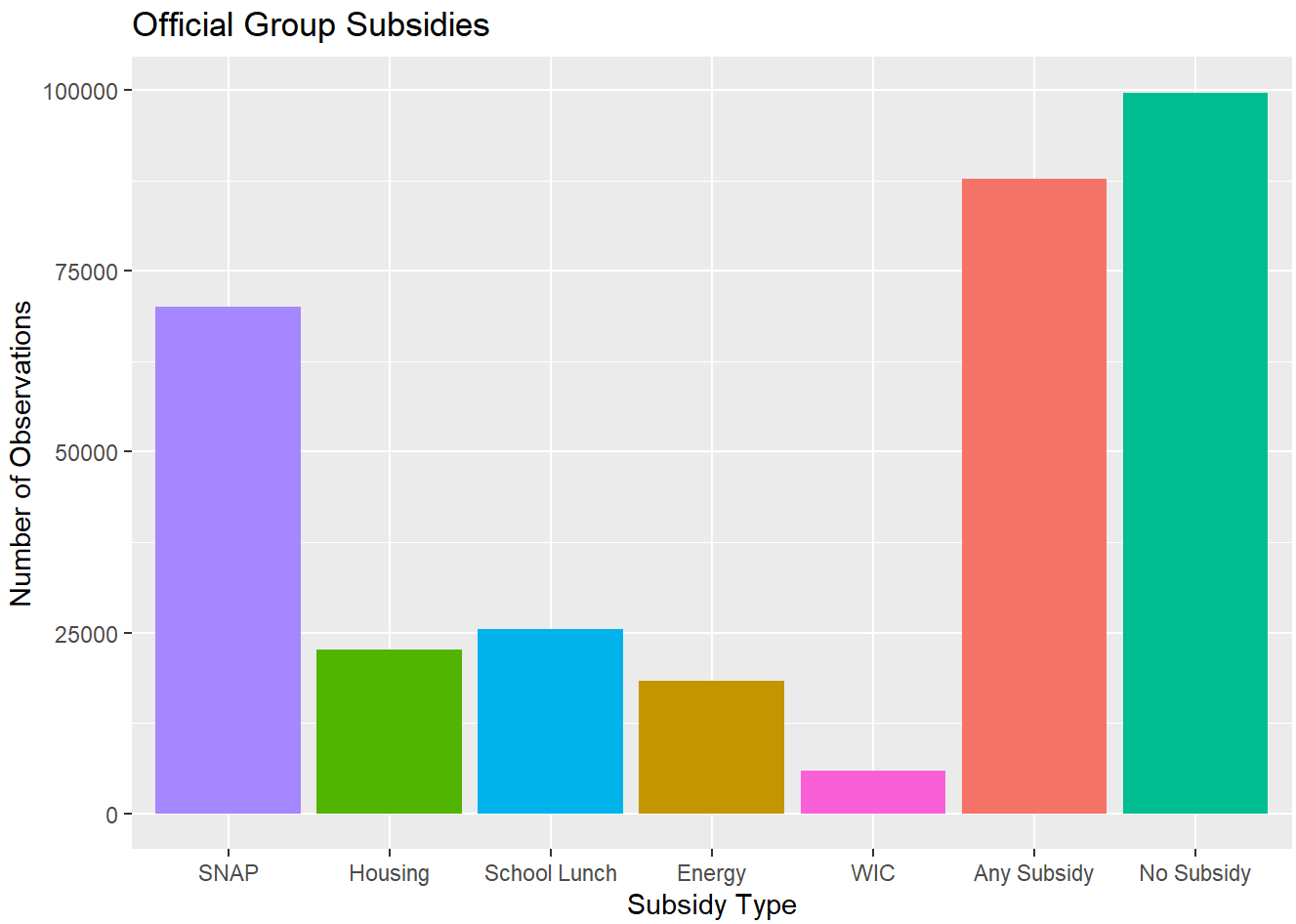
```
## [1] "Supplemental Group w/ Any Susbdiy: 96021"
```

```
print(paste("Supplemental Group Count of No Subsidies:", spm_no_sub))
```

```
## [1] "Supplemental Group Count of No Subsidies: 159042"
```

Next, we look at those individuals impoverished under the official definition.

```
# Filter to look at only those observations in the official poverty group.
off_group <- filter(df, off_pov==1)

# Count how many are not earning subsidies.
off_no_sub <- length(df$pov[(df$off_pov==1)&(df$snap == 0)&(df$house_sub == 0)&
                              (df$slunch == 0)&(df$wic == 0)&(df$energy == 0)])

# Count how many are earning subsidies.
off_w_sub <- length(df$pov[(df$off_pov==1)&((df$snap > 0)|
                                            (df$house_sub > 0)|
                                            (df$slunch > 0)|
                                            (df$wic > 0)|
                                            (df$energy > 0))])

# Count how many are earning subsidies of each type.
bar_vals <- c(length(off_group$snap[off_group$snap > 0]),
              length(off_group$house_sub[off_group$house_sub > 0]),
              length(off_group$slunch[off_group$slunch > 0]),
              length(off_group$energy[off_group$energy > 0]),
              length(off_group$wic[off_group$wic > 0]),
              off_w_sub, off_no_sub)

# Create our labels for the x-acis.
bar_labs <- c("SNAP", "Housing", "School Lunch", "Energy", "WIC",
              "Any Subsidy", "No Subsidy")

# Plot our data in a bar plot.
ggplot(mapping=aes(x=bar_labs, y=bar_vals, fill=bar_labs)) +
  geom_bar(stat="identity") +
  labs(title="Official Group Subsidies",
       x ="Subsidy Type",
       y = "Number of Observations")  +
  theme(legend.position = "none") +
  scale_x_discrete(limits = bar_labs)
```



We see that the difference between those not earning subsidies and those earning subsidies is reduced for this group. However, we still do have more individuals not earning subsidies.

Otherwise, this distribution is fairly similar to what we have been seeing. The biggest change is that the distributions of Housing and School Lunch subsidies are much closer.

```
# Print out bar chart values calculated earlier.
for (i in 1:5) {
  print(paste("Official Group Count of ", bar_labs[i], ": ", bar_vals[i],
              sep=""))
}
```

```
## [1] "Official Group Count of SNAP: 70119"
## [1] "Official Group Count of Housing: 22701"
## [1] "Official Group Count of School Lunch: 25466"
## [1] "Official Group Count of Energy: 18248"
## [1] "Official Group Count of WIC: 5840"
```

```
# Compare counts earning subs versus those not.
print(paste("Official Group w/ Any Susbdiy:", off_w_sub))
```

```
## [1] "Official Group w/ Any Susbdiy: 87796"
```

```
print(paste("Official Group Count of No Subsidies:", off_no_sub))
```

```
## [1] "Official Group Count of No Subsidies: 99701"
```

Total Subsidy Distribution

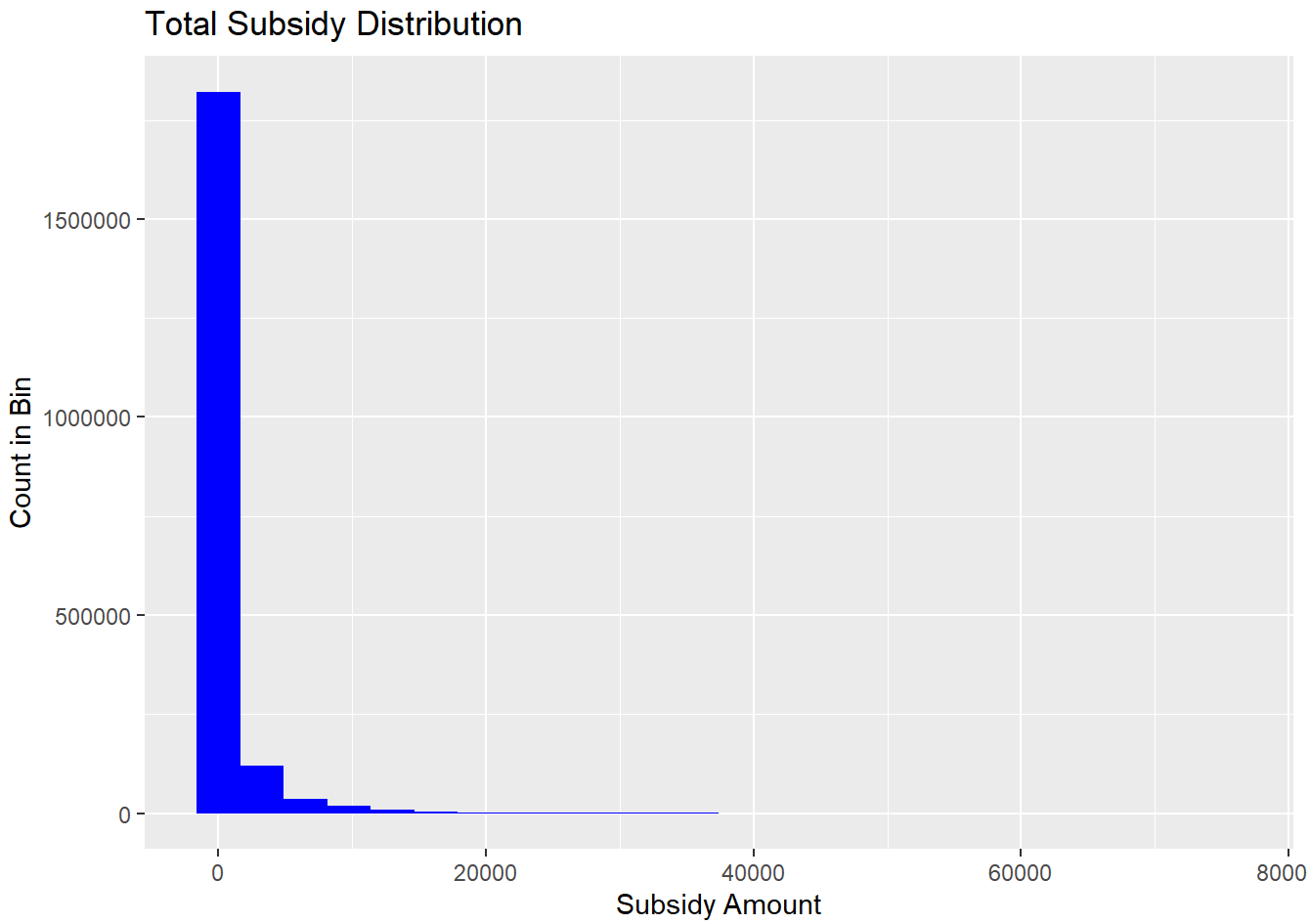
Next, we will use our new total subsidy variable to further investigate our subsidy variables.

```
summary(df$ttl_sub)   # Look at the box-and-whiskers distribution of ttl_sub
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     0.0   625.5     0.0 74819.0
```

We see that, overall, individuals are not earning subsidies in our data set.

```
# Create a histogram of the variable distribution.
ggplot(df, aes(x=ttl_sub)) +
  geom_histogram(bins=24, fill="blue") +
  labs(title="Total Subsidy Distribution",
       x ="Subsidy Amount",
       y = "Count in Bin")
```



Again, we see that most individuals are not earning that many subsidies. It is quite surprising that this suggests individuals may be earning up to $80,000 in total subsidies. Those observations could warrant some further investigation.

```
# What is the maximum amount of subsidies being earned?
print(max(df$ttl_sub))
```

```
## [1] 74819
```

```
# What is the second most?
print(max(df$ttl_sub[df$ttl_sub < max(df$ttl_sub)]))
```

```
## [1] 59928
```

We see that the maximum amount of subsidies an individual is earning is about 75000, and the second most is at about 60000. It could be interesting to see the full breakdown of subsidies these individuals are earning and compare that to their actual income.

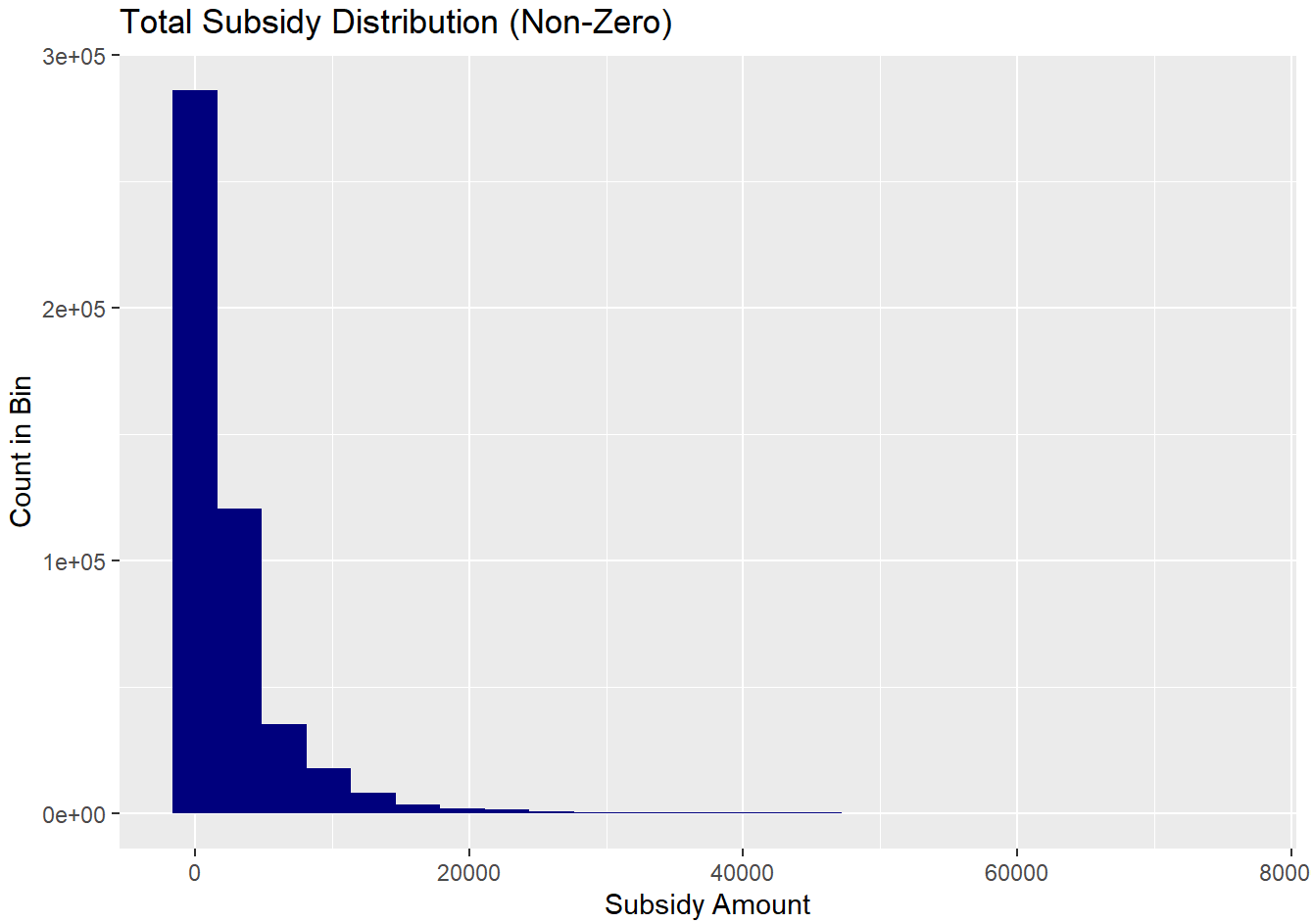
```
# Show the records of those observations earning the most subsidies.
filter(df, ttl_sub == max(df$ttl_sub))
```

```
##         X st off_pov age mar sex edu race hispanic agi hi_prem moop spm_pov
## 1 1786110 CA       1  41   M   M  HS    O        0   0       0   50       1
## 2 1786111 CA       1  36   M   F  HS    O        0   0       0 1050       1
## 3 1786112 CA       1  70   W   F <HS    O        0   0       0 1350       1
##   num_kid num_adlt mortgage spm_res spm_inc  snap house_sub slunch energy wic
## 1       5        3        R   78817    8000 20754     49884   3036    460 685
## 2       5        3        R   78817    8000 20754     49884   3036    460 685
## 3       5        3        R   78817    8000 20754     49884   3036    460 685
##   fed_tax tax_credit eitc fica st_tax cap_xpen wk_xpen cc_xpen spm_hi_prem
## 1       0          0    0 1130      0      372     372       0           0
## 2       0          0    0 1130      0      372     372       0           0
## 3       0          0    0 1130      0      372     372       0           0
##   med_xpen mc_pb cohabit ui_kids group  pov ttl_sub
## 1     2500     0       0       0  poor Both   74819
## 2     2500     0       0       0  poor Both   74819
## 3     2500     0       0       0  poor Both   74819
```

The fact that we have 3 observations here suggests we have a family of individuals. For our analysis, it might be useful to remove repeats like this. However, we would then need to ask which individual we would maintain? The oldest? Youngest? Male? Female? Any filtering rule we could apply would likely bias our overall results.

Moving on, the first bar chart is not the most useful since its scaling includes all of the observations who have not earned subsidies. A more usful chart would remove those so we can look at those individuals who are earning subsidies.

```
# We create a quick filter to remove anyone not earning subsidies.
regular_sub <- filter(df, (ttl_sub > 0))

# Recreate our last histogram with the new data.
ggplot(regular_sub, aes(x=ttl_sub)) +
  geom_histogram(bins=24, fill="navyblue") +
  labs(title="Total Subsidy Distribution (Non-Zero)",
       x ="Subsidy Amount",
       y = "Count in Bin")
```



This visualization shows us very similar data, but the scaling makes it more visually appealing now that we do not have the huge weight of observations who are not earning subsidies.

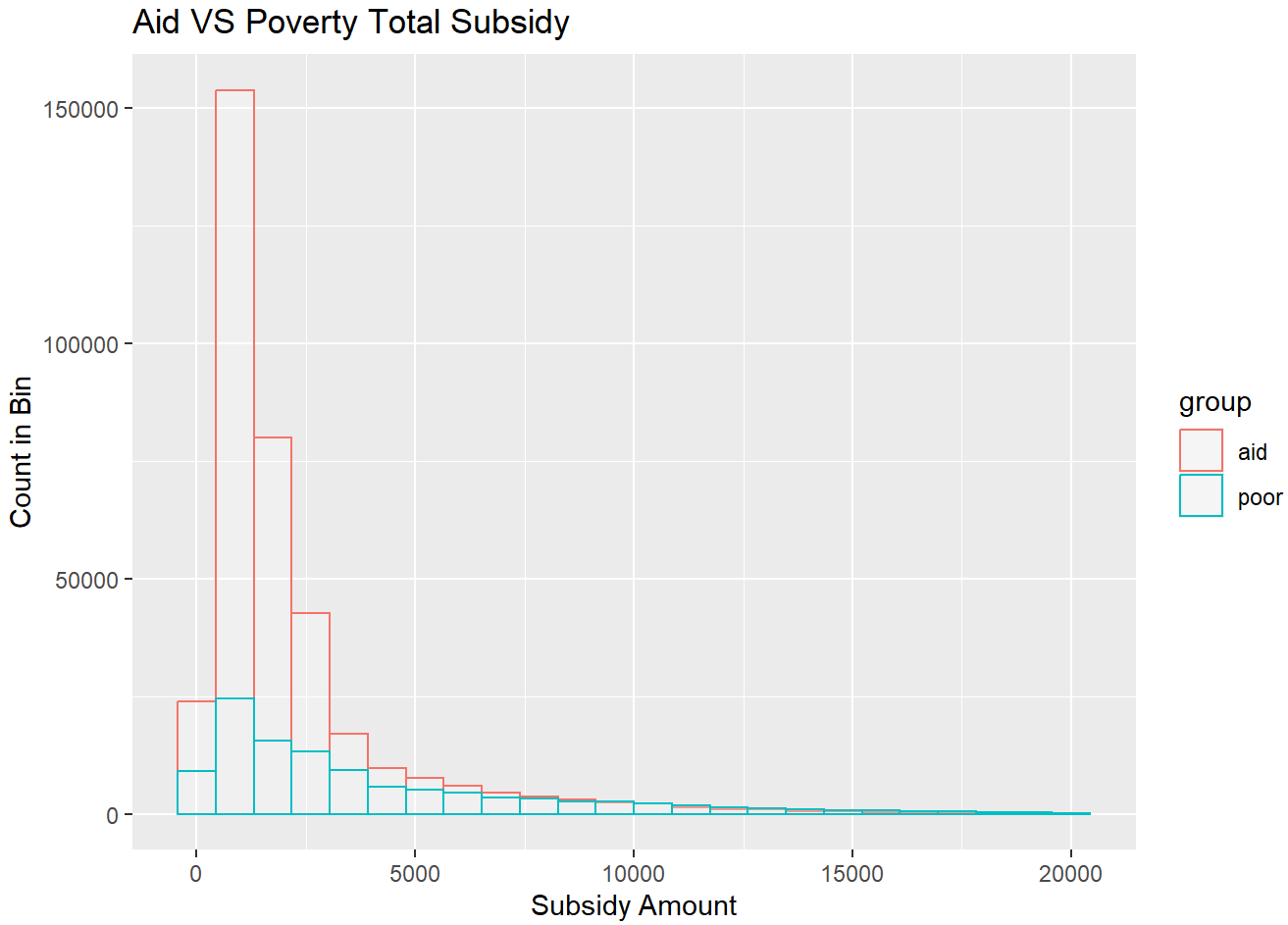
Multivariate Visualizations

Though we can uncover a lot about individual data distributions using univariate visualizations, multivariate visualizations can help us understand how each variable relates to our target variable, Group.

Subsidies By Group

We have already done quite a lot to examine how frequently individuals earn subsidies in our specific groups, but we have not looked at how much each group has been earning in subsidies. We would expect the poverty group to earn the most total subsidies, but that may not be the case.

```
# Create histogram comparing the total subsidies earned by the aid and poor
# groups. We filtered down to those earning less than 20,000 so we could focus
# on the more typical observations. We also removed observations earning no
# subsidies.
ggplot(filter(regular_sub, ttl_sub < 20000), aes(x=ttl_sub, color=group)) +
  geom_histogram(bins = 24, alpha=0.25, position="identity", fill="white") +
  labs(title="Aid VS Poverty Total Subsidy",
       x ="Subsidy Amount",
       y = "Count in Bin")
```



We see that, even though the Aid group is the larger group, especially when considering the number of observations actually earning subsidies, we see that on the right tail, we have more individuals in the poor group, suggesting that those individuals who are impoverished have more eligibility for these subsidies, which makes sense. We also see that most subsidy totals are under 5,000. The obvious question would be: Is this enough?

Next, we look at specific subsidy types.

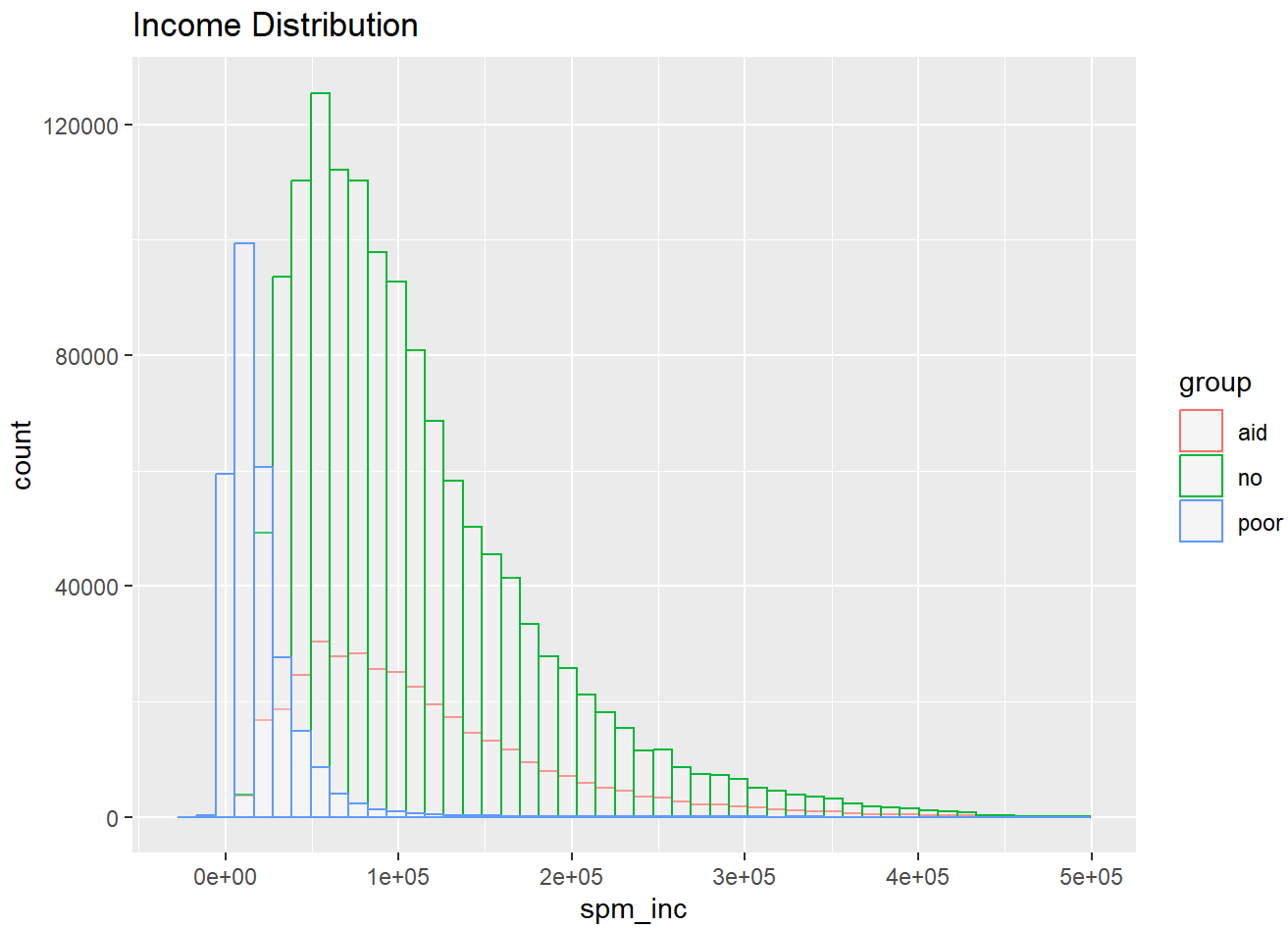
```
# Filter to those earning SNAP subsidies.
sub_df <- filter(df, (snap > 0))
# Create a plot of the distribution.
g1 <- ggplot(sub_df, aes(x=snap, color=group)) +
  geom_histogram(bins=24, alpha=0.25, position="identity", fill="white") +
  labs(title="SNAP", x ="Subsidy Amount", y = "Count in Bin") +
  theme(legend.position = "none")   # We'll use 1 Legend on the last plot.

# Filter to those earning school lunch subsidies.
sub_df <- filter(df, (slunch > 0))
g2 <- ggplot(sub_df, aes(x=slunch, color=group)) +
  geom_histogram(bins=24, alpha=0.25, position="identity", fill="white") +
  labs(title="School Lunch", x ="Subsidy Amount", y = "Count in Bin") +
  theme(legend.position = "none")

# Filter to those individuals earning housing subsidies.
sub_df <- filter(df, (house_sub > 0))
g3 <- ggplot(sub_df, aes(x=house_sub, color=group)) +
  geom_histogram(bins=24, alpha=0.25, position="identity", fill="white") +
  labs(title="Housing", x ="Subsidy Amount", y = "Count in Bin") +
  theme(legend.position = "none")

# Filter to those individuals earning energy subsidies.
sub_df <- filter(df, (energy > 0))
g4 <- ggplot(sub_df, aes(x=energy, color=group)) +
  geom_histogram(bins=24, alpha=0.25, position="identity", fill="white") +
  labs(title="Energy", x ="Subsidy Amount", y = "Count in Bin") +
  theme(legend.position = "none")

# Filter to those individuals earning WIC subsidies.
sub_df <- filter(df, (wic > 0))
g5 <- ggplot(sub_df, aes(x=wic, color=group)) +
  geom_histogram(bins=24, alpha=0.25, position="identity", fill="white") +
  labs(title="WIC", x ="Subsidy Amount", y = "Count in Bin") +
  theme(legend.position = c(1.5, 0.5))

# Create plot of all histograms as defined above.
grid.arrange(g1, g2, g3, g4, g5, nrow=2, top="Subsidy Distributions")
```

We see that SNAP, Energy, and WIC subsidies are reasonably close between the two groups. However, housing subsidies seem to greatly favor the Poor group, and School Lunch subsidies seem to favor the aid group. This seems to suggest at least one difference between the needs of these groups.

Income Categories by Group

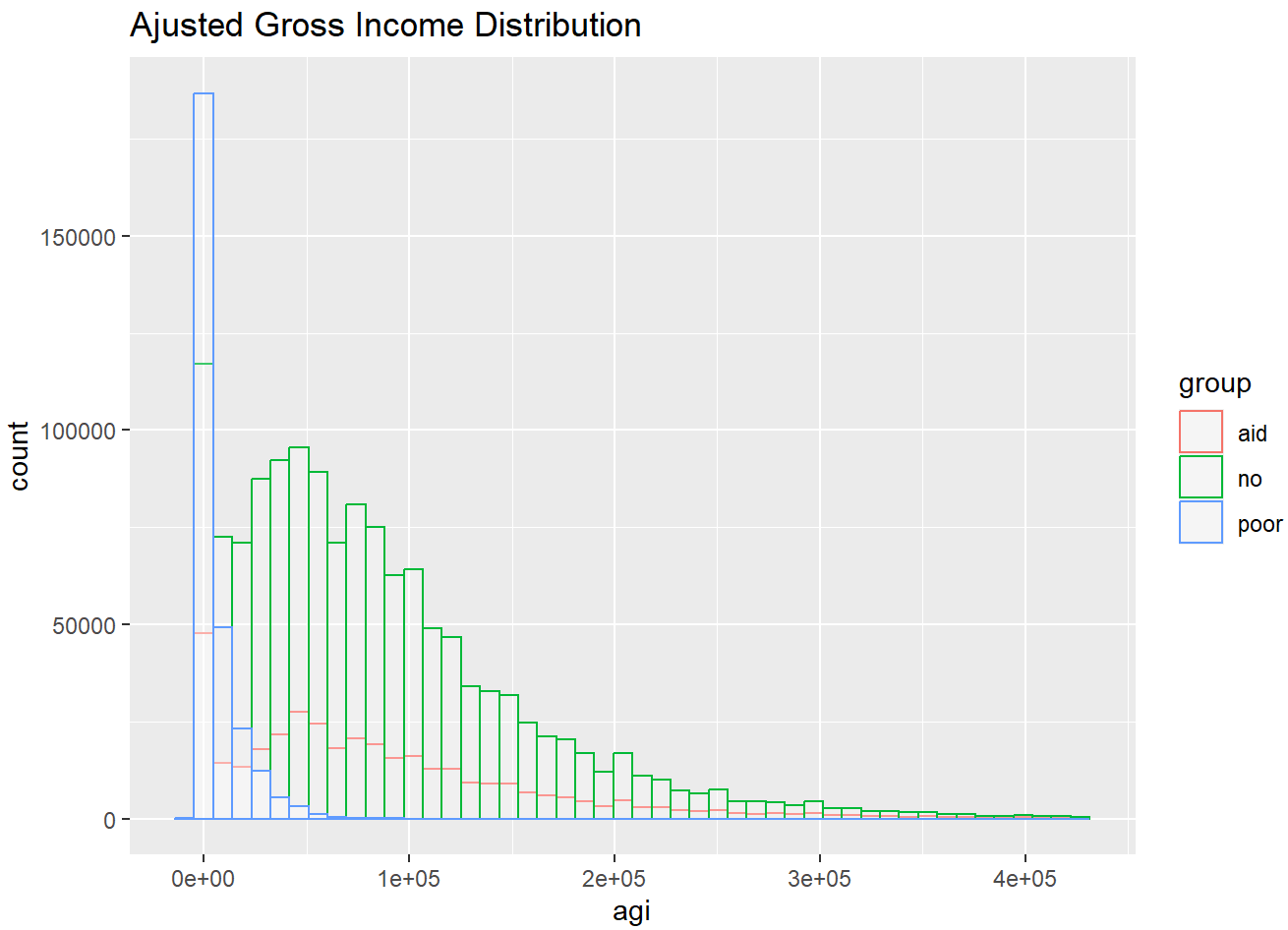
Next, we can examine our income categories.

```
# Create a histogram of Income by Group.
ggplot(df, aes(x=spm_inc, color=group)) +
  geom_histogram(bins=48, alpha=0.25, position="identity", fill="white") +
  labs(title="Income Distribution", xlab = "Income", ylab="Count")
```



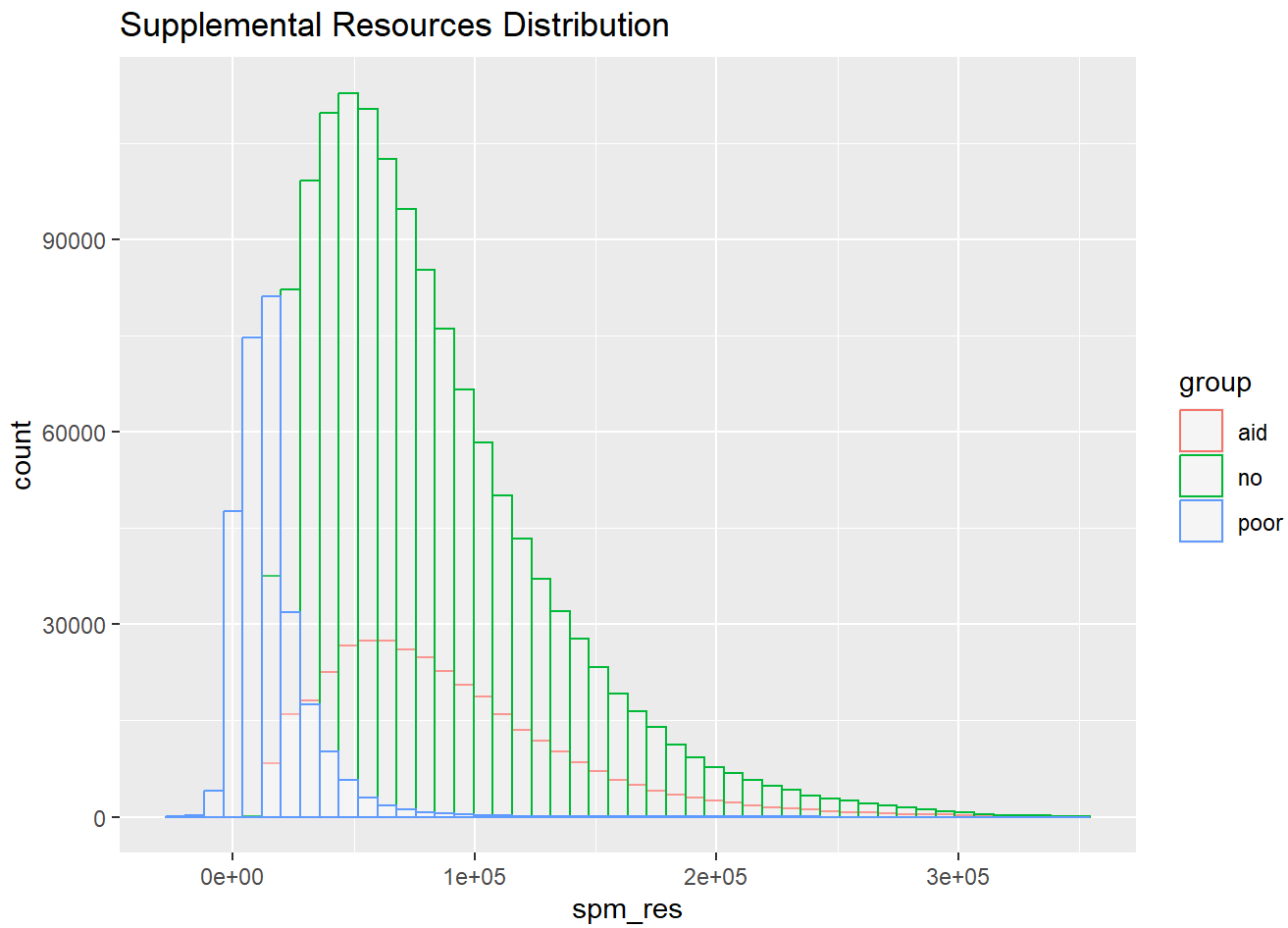
In this initial plot, we see that the distribution of income is very similar between the aid and the no groups. The poor group seems to be much tighter distributed, with most observations below $50,000.

```
# Create a histogram of AGI distributions between groups.
ggplot(df, aes(x=agi, color=group)) +
  geom_histogram(bins=48, alpha=0.25, position="identity", fill="white") +
  labs(title="Ajusted Gross Income Distribution", xlab = "AGI", ylab="Count")
```



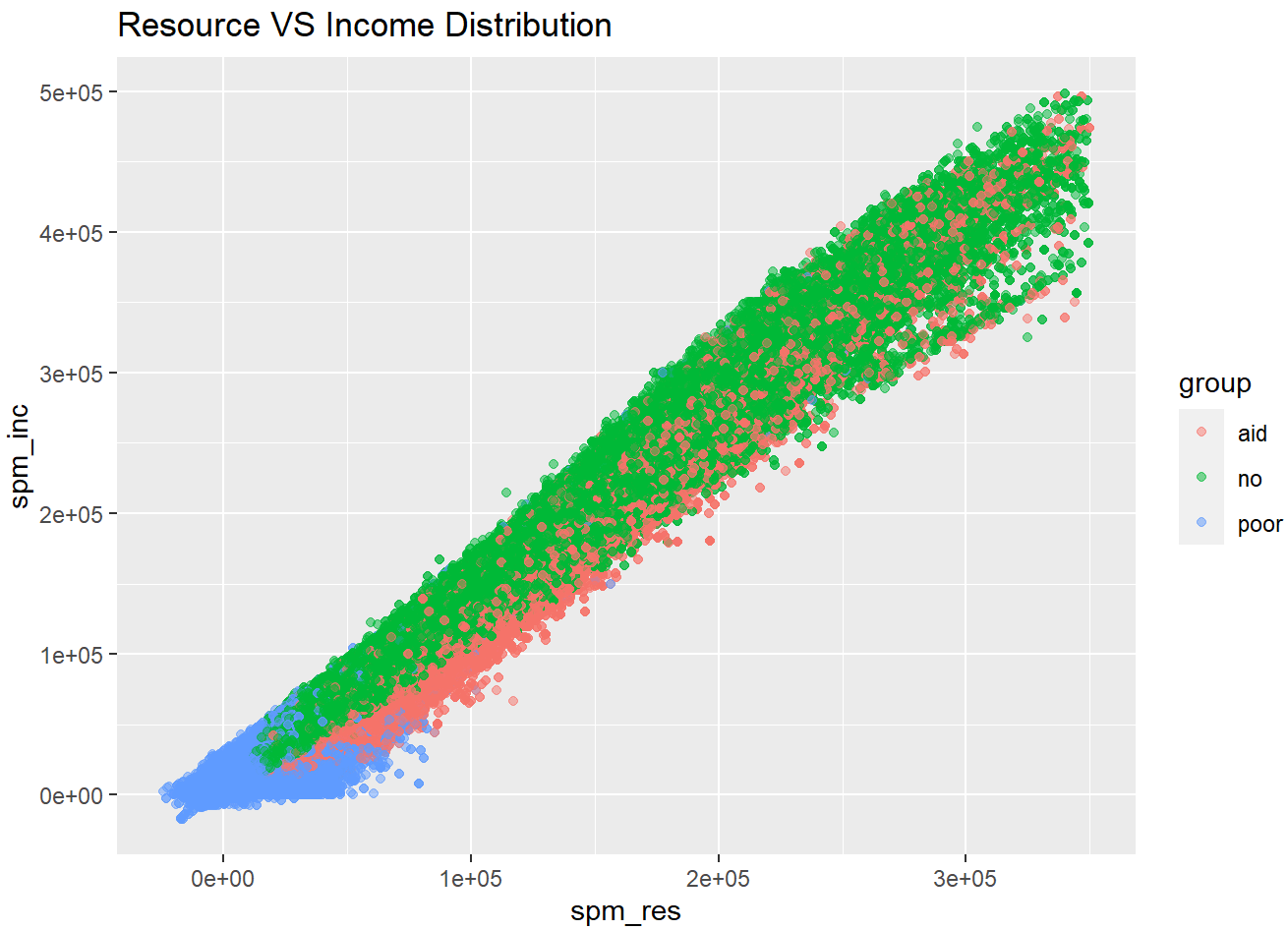
Our results for AGI are very similar to the results we found for Income. We see even more skewing in the poverty group, pushing most values for that group to 0.

```
# Create a histogram for total Resource by Group.
ggplot(df, aes(x=spm_res, color=group)) +
  geom_histogram(bins=48, alpha=0.25, position="identity", fill="white")+
  labs(title="Supplemental Resources Distribution",
       xlab = "SPM Resource", ylab="Count")
```



This figure is very similar to what we saw for Income. It seems that there is slightly less skewing for the aid and no groups. Generally, this seems to be a very similar variable to Income.

```
# Create a histogram for total Resource by Group.
ggplot(df, aes(x=spm_res, y=spm_inc, color=group)) +
  geom_point(alpha=0.5) +
  labs(title="Resource VS Income Distribution",
       xlab = "SPM Resource", ylab="SPM Income")
```



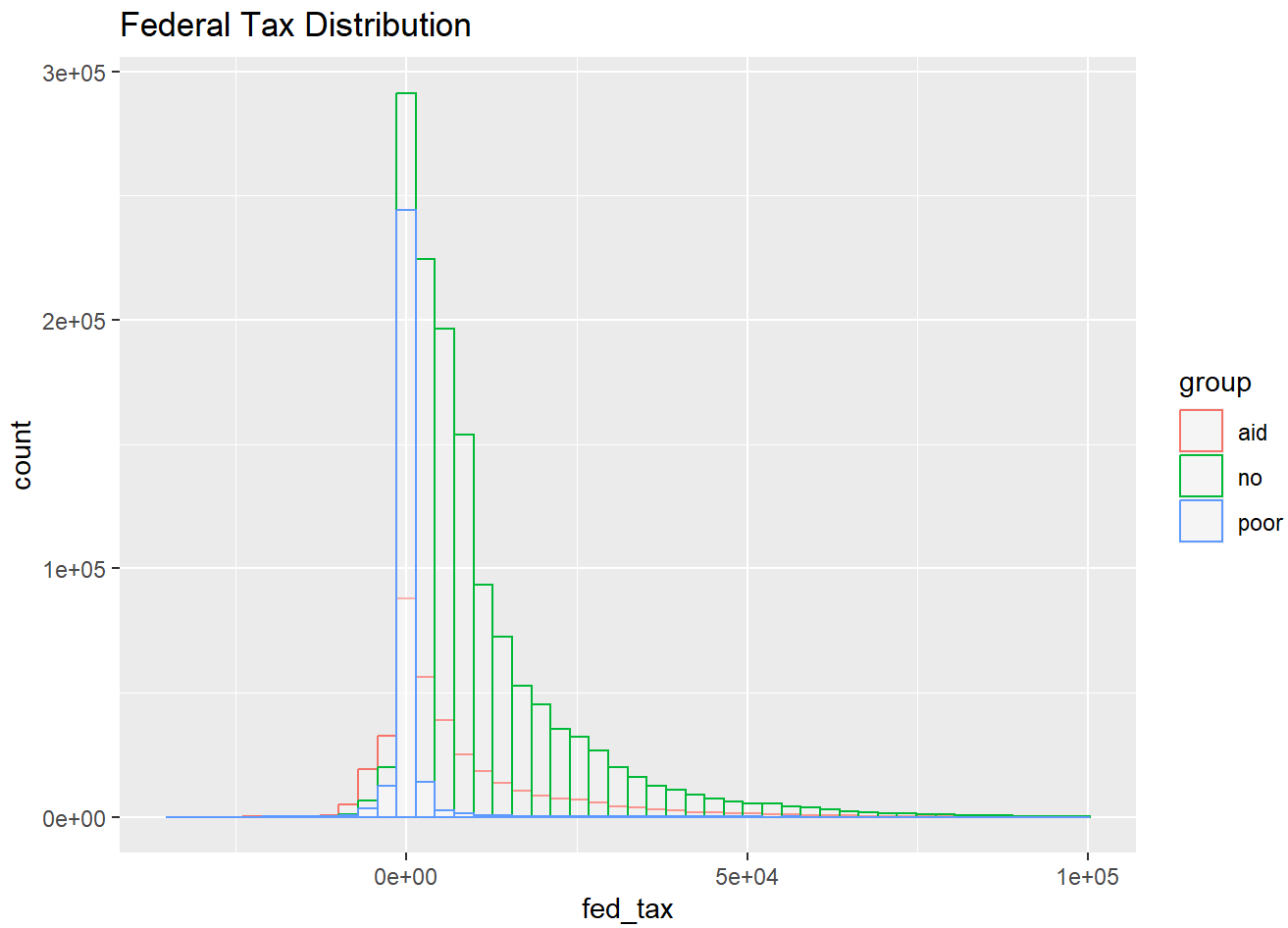
We see a very tight linear relationship between resource and income here. Income seems to have slightly higher values. And for our groups, we see that the poverty group reaches the lowest values, whereas the no group seems to have a slight preference for higher values of income. However, we do see a few poverty group observations at higher income and resource values. It might be interesting to investigate those higher income observations in that group.

Generally, this plot seems to suggest more similarity between the aid group and the no group.

Tax Categories by Group

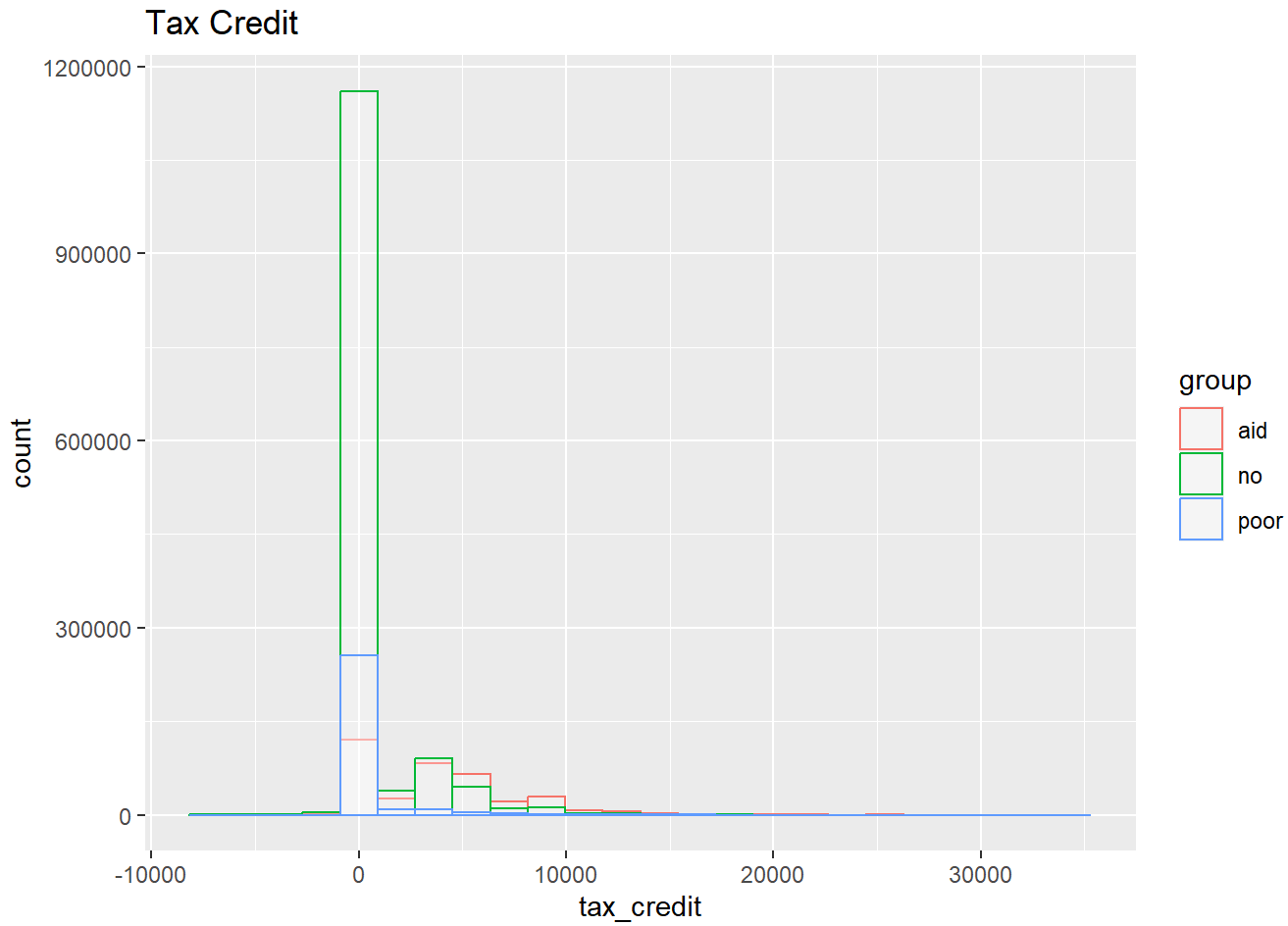
Next, we can examine the tax payments made by each group.

```
# Create a histogram of federal taxes due by group..
ggplot(df, aes(x=fed_tax, color=group)) +
  geom_histogram(bins=48, alpha=0.25, position="identity", fill="white") +
  labs(title="Federal Tax Distribution", xlab = "Fed Tax", ylab="Count")
```



The distributions of our variables here are actually fairly similar. We see that the poor group has a much tighter distribution around 0, but all of our distributions share that center and see some right skewing. Interestingly, the aid group has the most negative tax values.

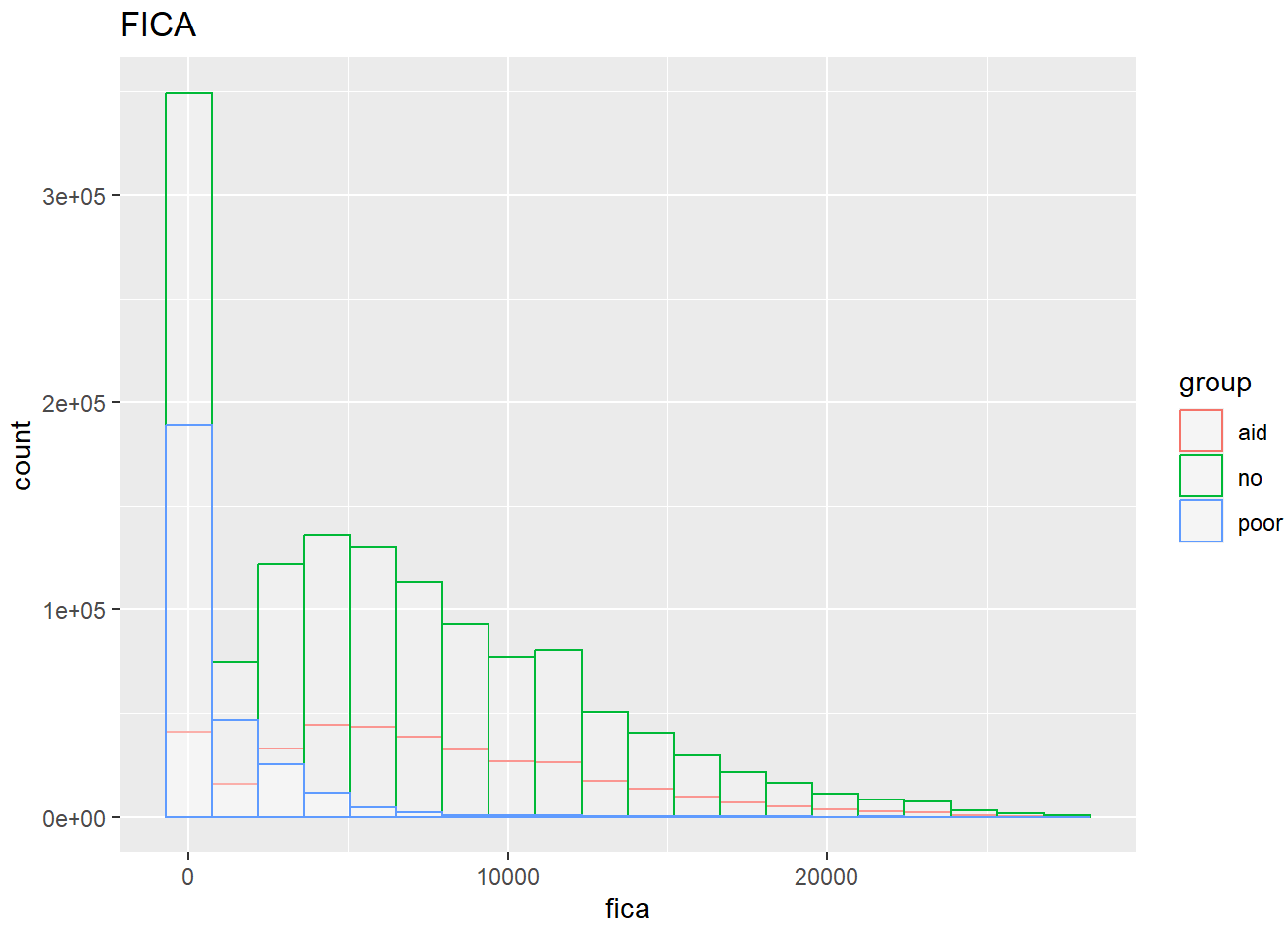
```
# Create a histogram of tax credits earned by group.
g1 <- ggplot(df, aes(x=tax_credit, color=group)) +
  geom_histogram(bins=24, alpha=0.25, position="identity", fill="white") +
  labs(title="Tax Credit", xlab = "Tax Credit", ylab="Count")
g1   # We saved this plot for a later visualization.
```



Our distribution of tax credits is very strange. Though the shape is odd, we see that the aid group has the highest relative percent of individuals earning tax credits. And somehow, we have values who earned negative tax credit. Interestingly, the poor group does not seem to earn that many tax credits. Our programs likely do not focus on providing credits to these individuals since they are likely already paying little to no money in actual taxes.

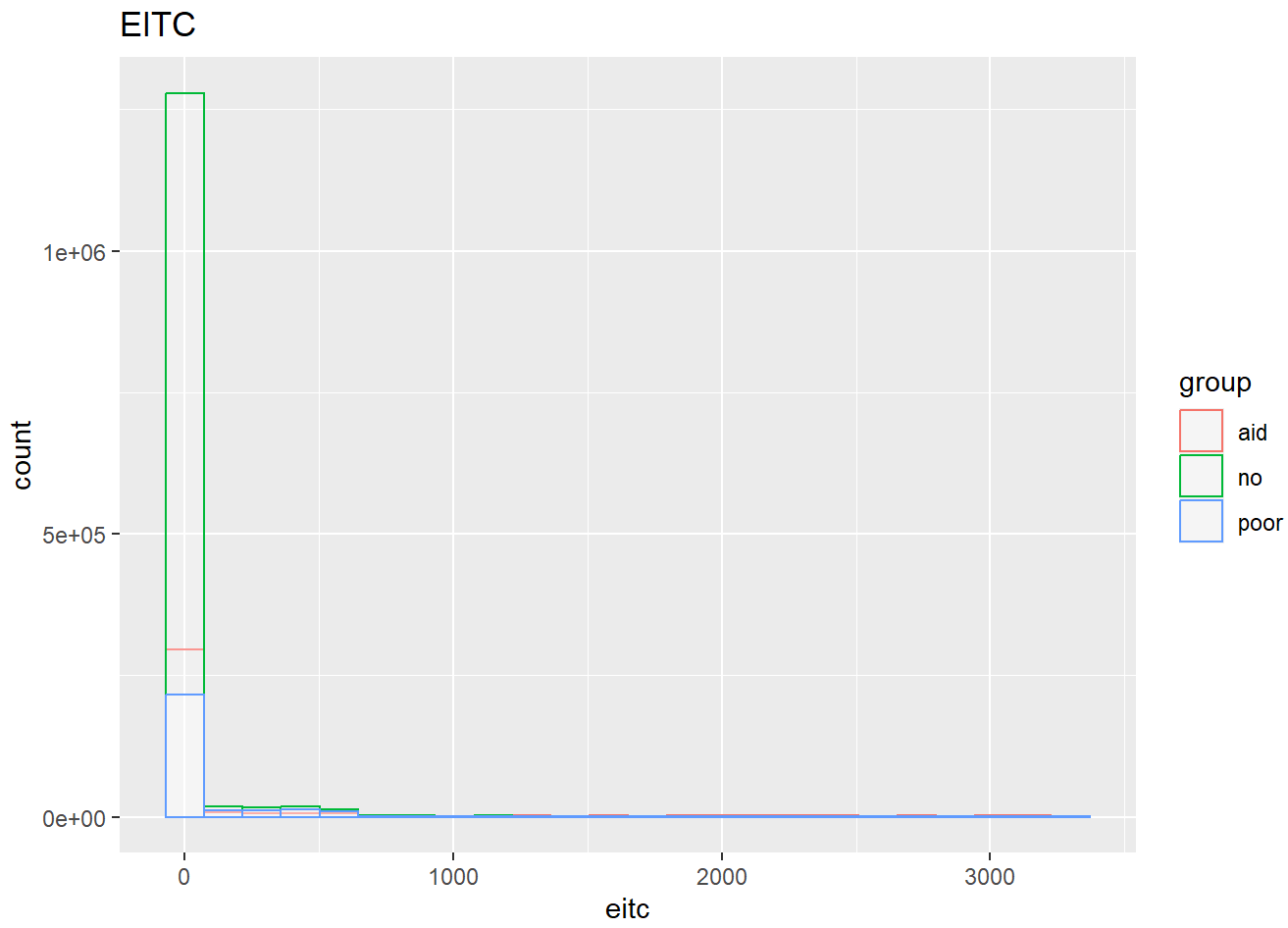
Generally, this distribution seems discreet. For credit calculations, you sum the values earned for select credit categories, which is not typically adjusted by relative income.

```
# Create a histogram of FICA credit earned by group.
g2 <-ggplot(df, aes(x=fica, color=group)) +
  geom_histogram(bins=20, alpha=0.25, position="identity", fill="white") +
  labs(title="FICA", xlab = "FICA", ylab="Count")
g2
```

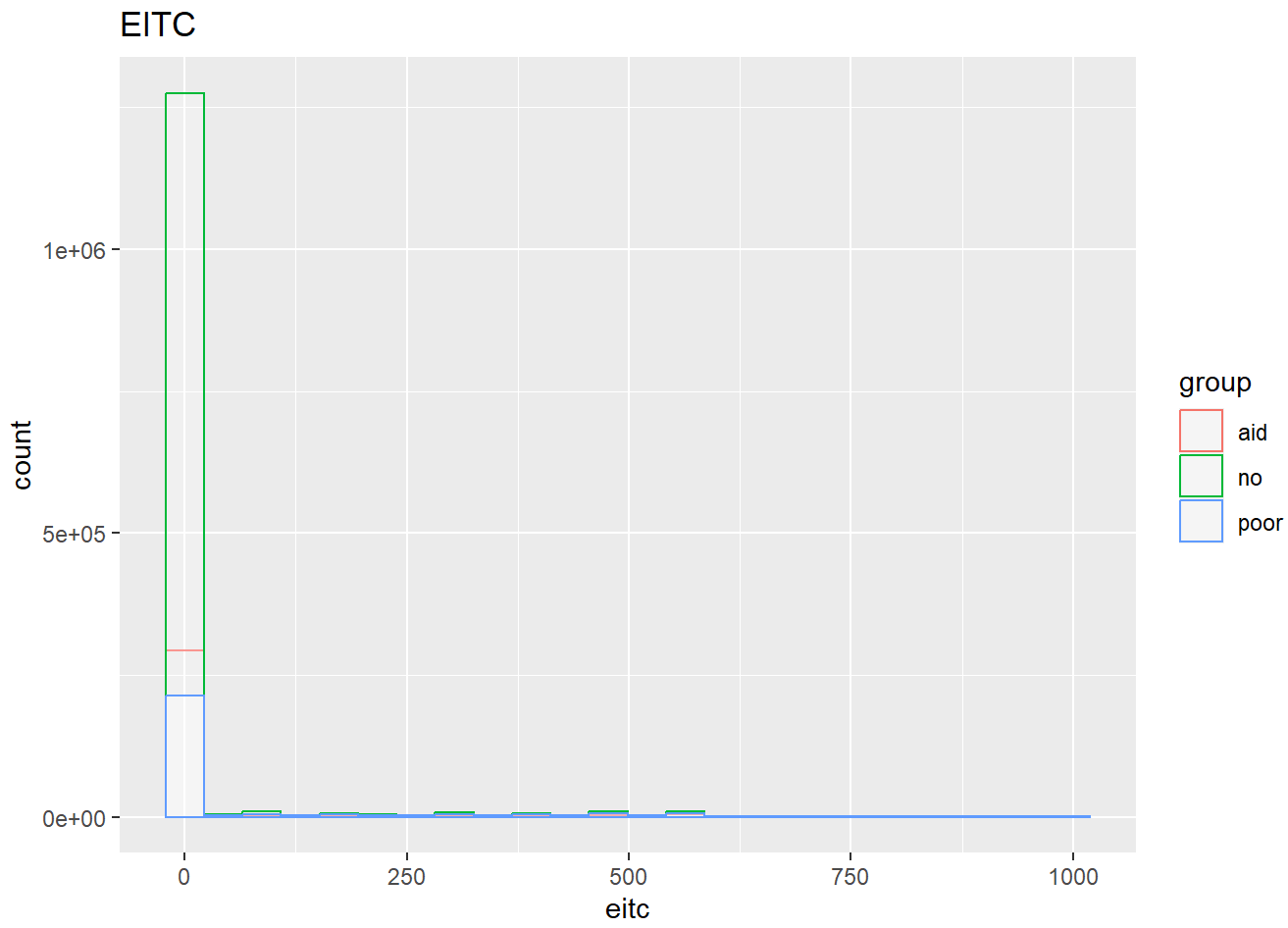


For our FICA distribution, we seem to return to the trends of the income and resource variables.

```
# Create a histogram of EITC by group.
g3 <- ggplot(df, aes(x=eitc, color=group)) +
  geom_histogram(bins=24, alpha=0.25, position="identity", fill="white") +
  labs(title="EITC", xlab = "EITC", ylab="Count")
g3
```



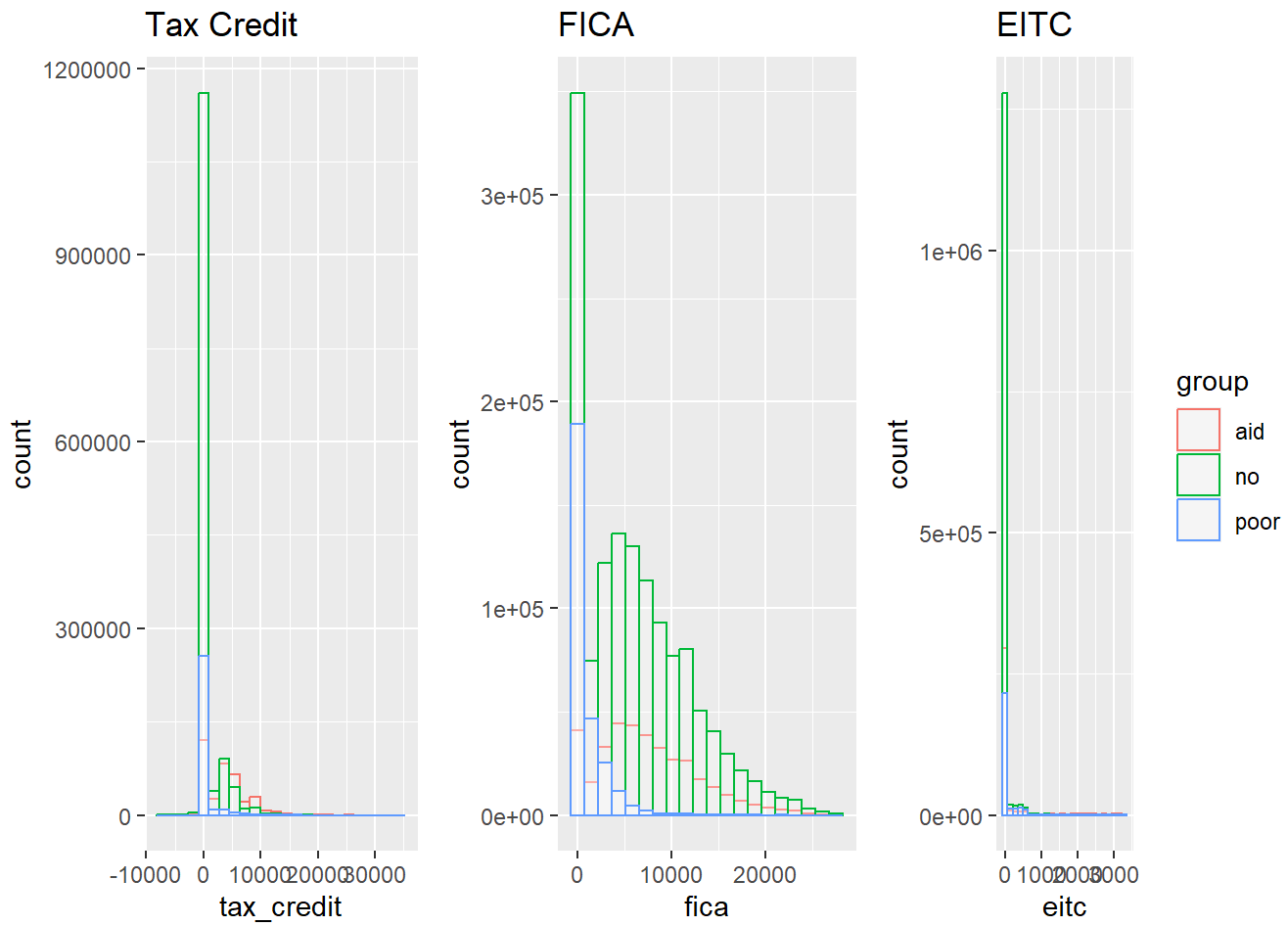
Our EITC credits seem to be much less common than other credit types. At higher values, we see that mostly aid individuals are still earning. It might help to focus on individuals earning $1,000 or less to zoom in on the distribution.



Even zoomed in, it does not appear that many individuals earn EITC credits in our distribution. Since we lack non-zero values, it might help to remove this variable from our data set because it could be biased in our models.

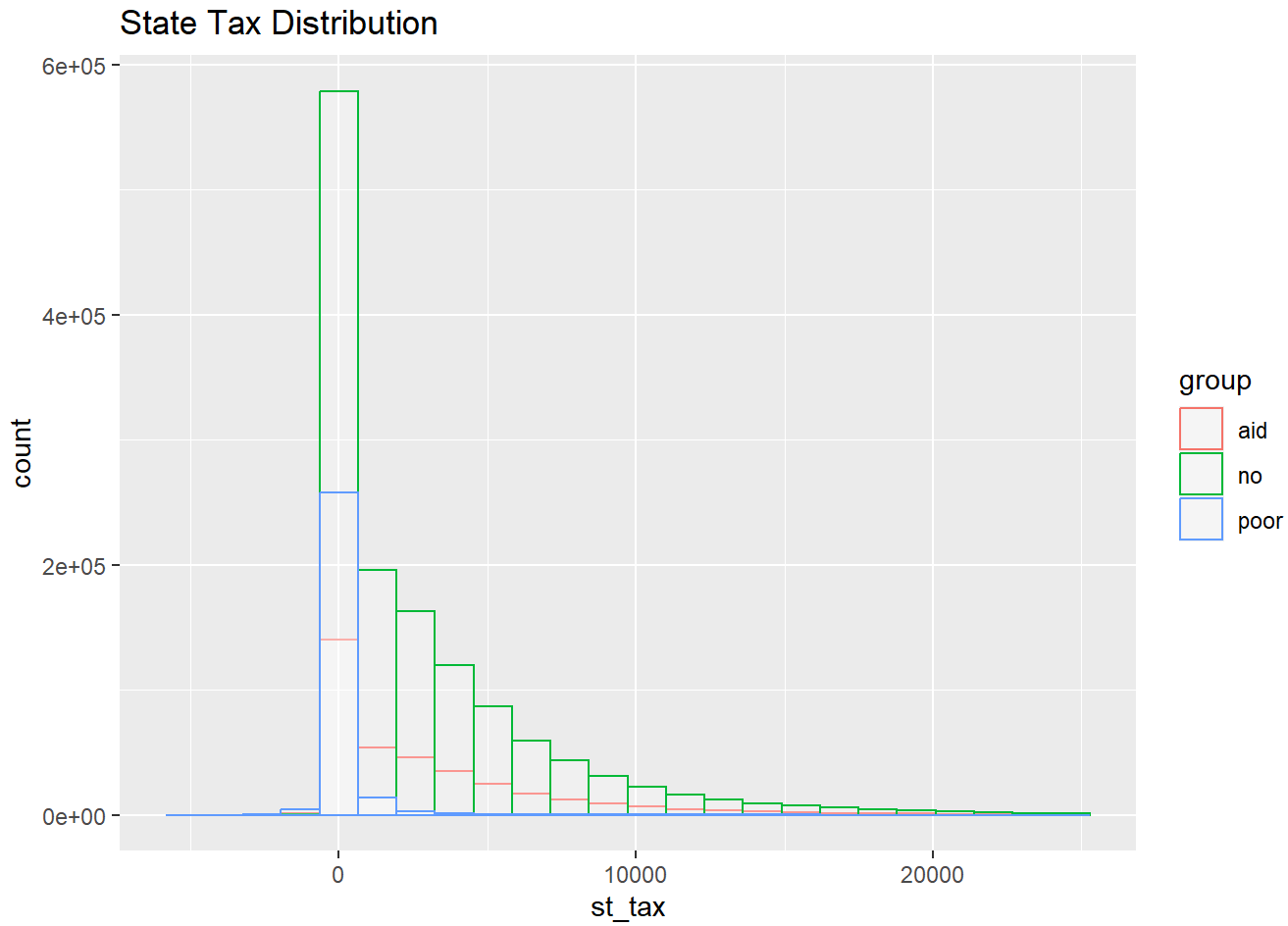
Next, we combine our graphs for convenience of reporting.

```
# Combine our visualizations into 1 graph.
grid.arrange(g1 + theme(legend.position = "none"), g2 + theme(legend.position = "none"), g3, nro
w=1)
```



For state tax, it will be interesting to see if different state laws change the distributions of our variables.

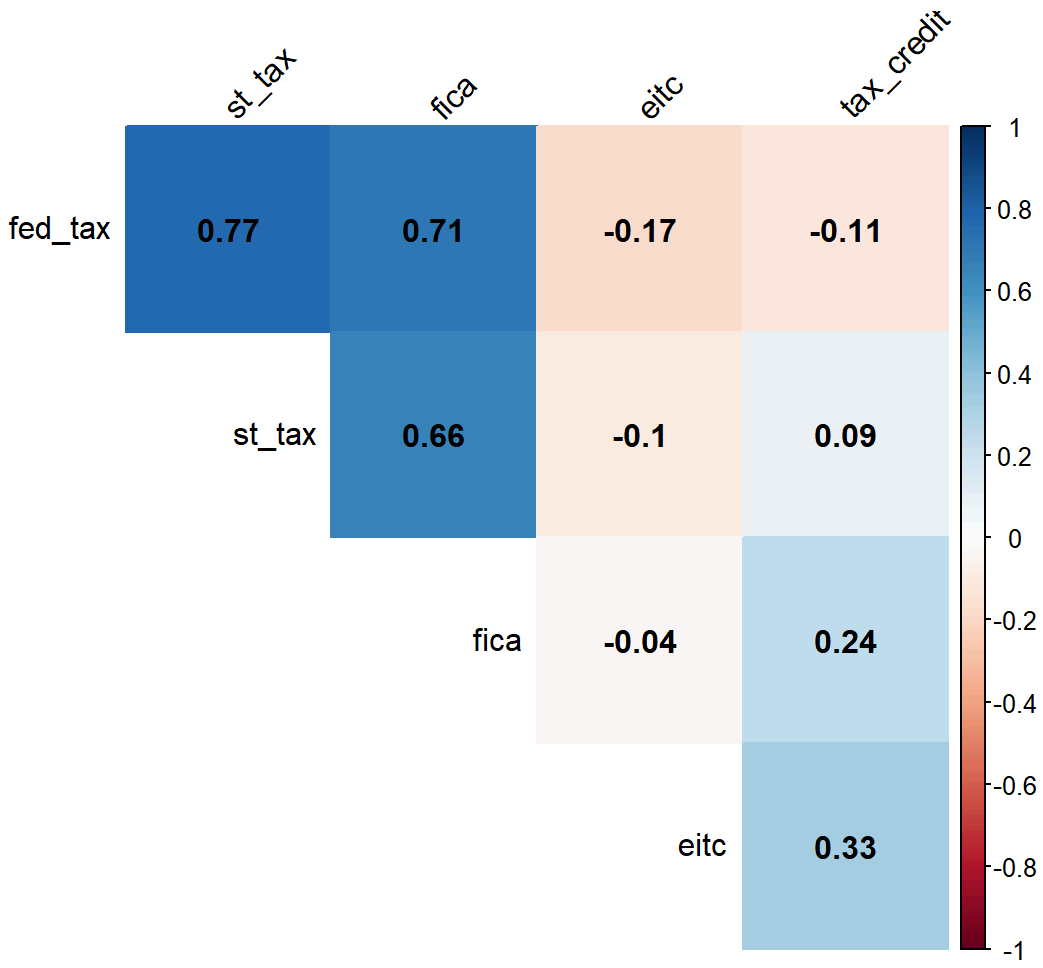
```
ggplot(df, aes(x=st_tax, color=group)) +
  geom_histogram(bins=24, alpha=0.25, position="identity", fill="white") +
  labs(title="State Tax Distribution", xlab = "State Tax", ylab="Count")
```



Interestingly, we now see all groups following the same approximately 0 Median, right skewed distribution. Some states do not have their own tax, which likely contributes to this distribution.

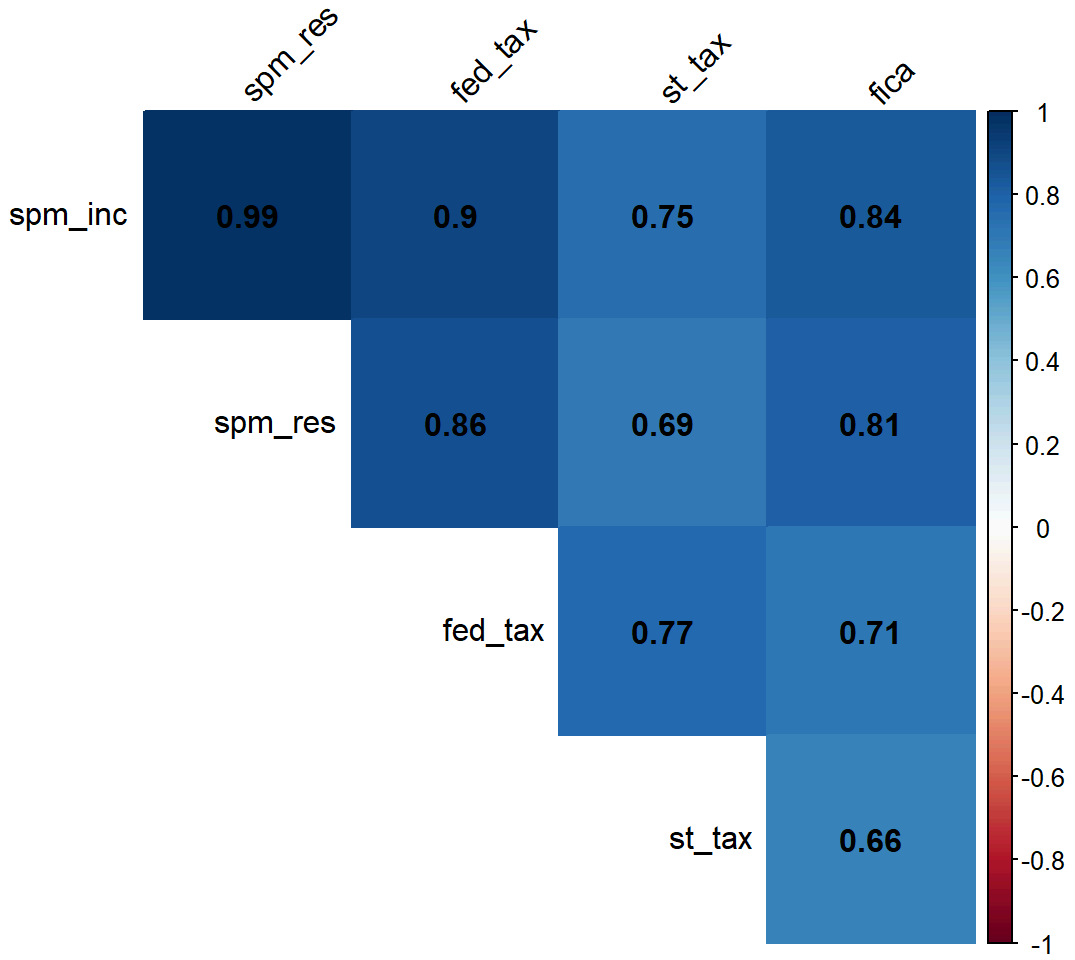
Next we look at the correlations between our tax variables.

```
# Create a correlation plot between our tax variables.
corrplot(cor(select(df, fed_tax, st_tax, fica, eitc, tax_credit)),
         method="color", type="upper", diag=FALSE,
         addCoef.col = "black", tl.col="black", tl.srt=45)
```



We see the strongest correlations between state tax, federal tax, and fica. This makes sense because they are very income-centric metrics. It might help to look at the correlation between these and income/resource.

```
# Create a correlation plot between our income and primary tax variables.
corrplot(cor(select(df, spm_inc, spm_res, fed_tax, st_tax, fica)),
         method="color", type="upper", diag=FALSE,
         addCoef.col = "black", tl.col="black", tl.srt=45)
```



We see very strong correlations here, which could lead to multicolinearity in our models. It may help to remove one of spm_res or spm_inc due to their extremely strong correlation. If we were to make such a removal, it would likely be better to keep spm_res because it has slightly lower correlations compared to our other variables.

Demographic Variables

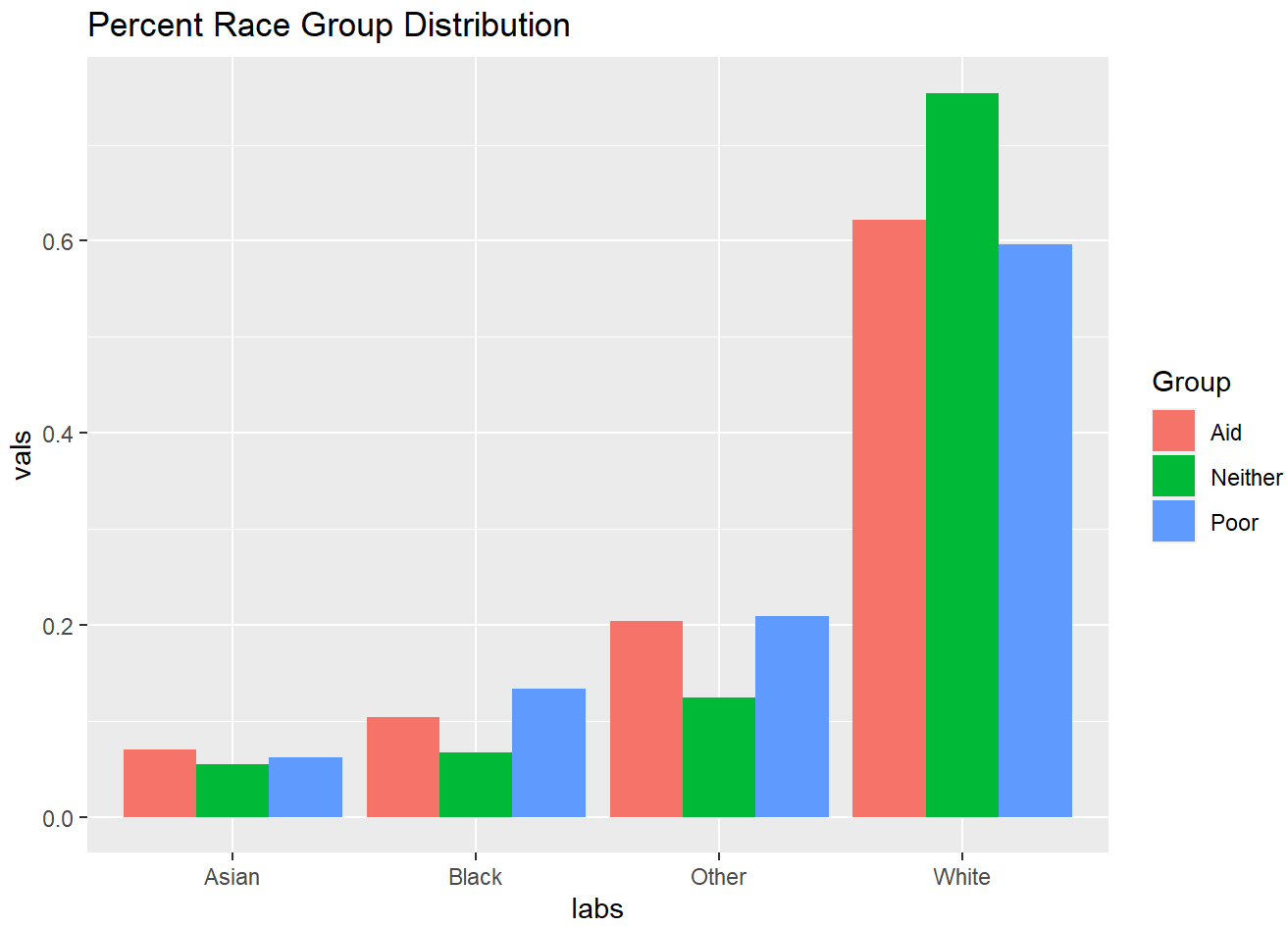
Next, we see if there are any demographic differences between our groups.

```
# Loop through each target group.
for (i in c("poor", "aid", "no")) {
  # Get a total length to determine percents with.
  y <- length(df$group[df$group == i])

  # Find percents of each race group present in data set.
  x1 <- length(df$group[(df$group == i)&(df$race == "A")])
  x2 <- length(df$group[(df$group == i)&(df$race == "B")])
  x3 <- length(df$group[(df$group == i)&(df$race == "O")])
  x4 <- length(df$group[(df$group == i)&(df$race == "W")])

  # Print out our results.
  print(paste("Group", i))
  print(paste("   Percent White:", x4/y))
  print(paste("   Percent Black:", x2/y))
  print(paste("   Percent Asian:", x1/y))
  print(paste("   Percent Other:", x3/y))
}
```

```
## [1] "Group poor"
## [1] "   Percent White: 0.595805799870041"
## [1] "   Percent Black: 0.133204557801631"
## [1] "   Percent Asian: 0.0615601487073327"
## [1] "   Percent Other: 0.209429493620995"
## [1] "Group aid"
## [1] "   Percent White: 0.622033240371882"
## [1] "   Percent Black: 0.10414252009041"
## [1] "   Percent Asian: 0.0703735618735082"
## [1] "   Percent Other: 0.203450677664199"
## [1] "Group no"
## [1] "   Percent White: 0.754070484549251"
## [1] "   Percent Black: 0.0668333193032797"
## [1] "   Percent Asian: 0.0551505546751189"
## [1] "   Percent Other: 0.12394564147235"
```

First, we look at the differences in race distribution from a quantitative perspective. We see that the no group has the highest percent of White individuals. The Poor and Aid group otherwise have much more similar distributions, with the Poor group holding a slight preference to Black individuals and the aid group holding a slight preference to Asian individuals.

```
# Here, I copied the above code and formally defined it for each group.
# Poor group calculations.
i <- "poor"
yp <- length(df$group[df$group == i])
x1p <- length(df$group[(df$group == i)&(df$race == "A")])/yp
x2p <- length(df$group[(df$group == i)&(df$race == "B")])/yp
x3p <- length(df$group[(df$group == i)&(df$race == "O")])/yp
x4p <- length(df$group[(df$group == i)&(df$race == "W")])/yp

# Aid group calculatins.
i <- "aid"
ya <- length(df$group[df$group == i])
x1a <- length(df$group[(df$group == i)&(df$race == "A")])/ya
x2a <- length(df$group[(df$group == i)&(df$race == "B")])/ya
x3a <- length(df$group[(df$group == i)&(df$race == "O")])/ya
x4a <- length(df$group[(df$group == i)&(df$race == "W")])/ya

# No group calculations.
i <- "no"
yn <- length(df$group[df$group == i])
x1n <- length(df$group[(df$group == i)&(df$race == "A")])/yn
x2n <- length(df$group[(df$group == i)&(df$race == "B")])/yn
x3n <- length(df$group[(df$group == i)&(df$race == "O")])/yn
x4n <- length(df$group[(df$group == i)&(df$race == "W")])/yn

# Assign bar chart values from the above calculations.
bar_vals <- c(x1p, x1a, x1n, x2p, x2a, x2n, x3p, x3a, x3n, x4p, x4a, x4n)

# Create labels for our bar chart.
# We have one set of labels for our Race category...
bar_labs <- c("Asian", "Asian", "Asian", "Black", "Black", "Black", "Other",
              "Other", "Other", "White", "White", "White")
# ...And another set of labels for our Group category.
bar_grps <- c("Poor", "Aid", "Neither", "Poor", "Aid", "Neither", "Poor",
              "Aid", "Neither", "Poor", "Aid", "Neither")

# Create a data frame out of these values to create our viz with.
graph_df <- data.frame("vals" = bar_vals, "labs" = bar_labs, "Group" = bar_grps)

# Define a visualization using the above data frame.
gr <- ggplot(graph_df, aes(x=labs, y=vals, fill=Group)) +
  geom_bar(stat="identity", position=position_dodge(preserve="single")) +
  labs(title = "Percent Race Group Distribution",
       xlab = "Race", ylab = "Percent Distribution")

gr
```



Visually, we see just how different the No group is from the other two.

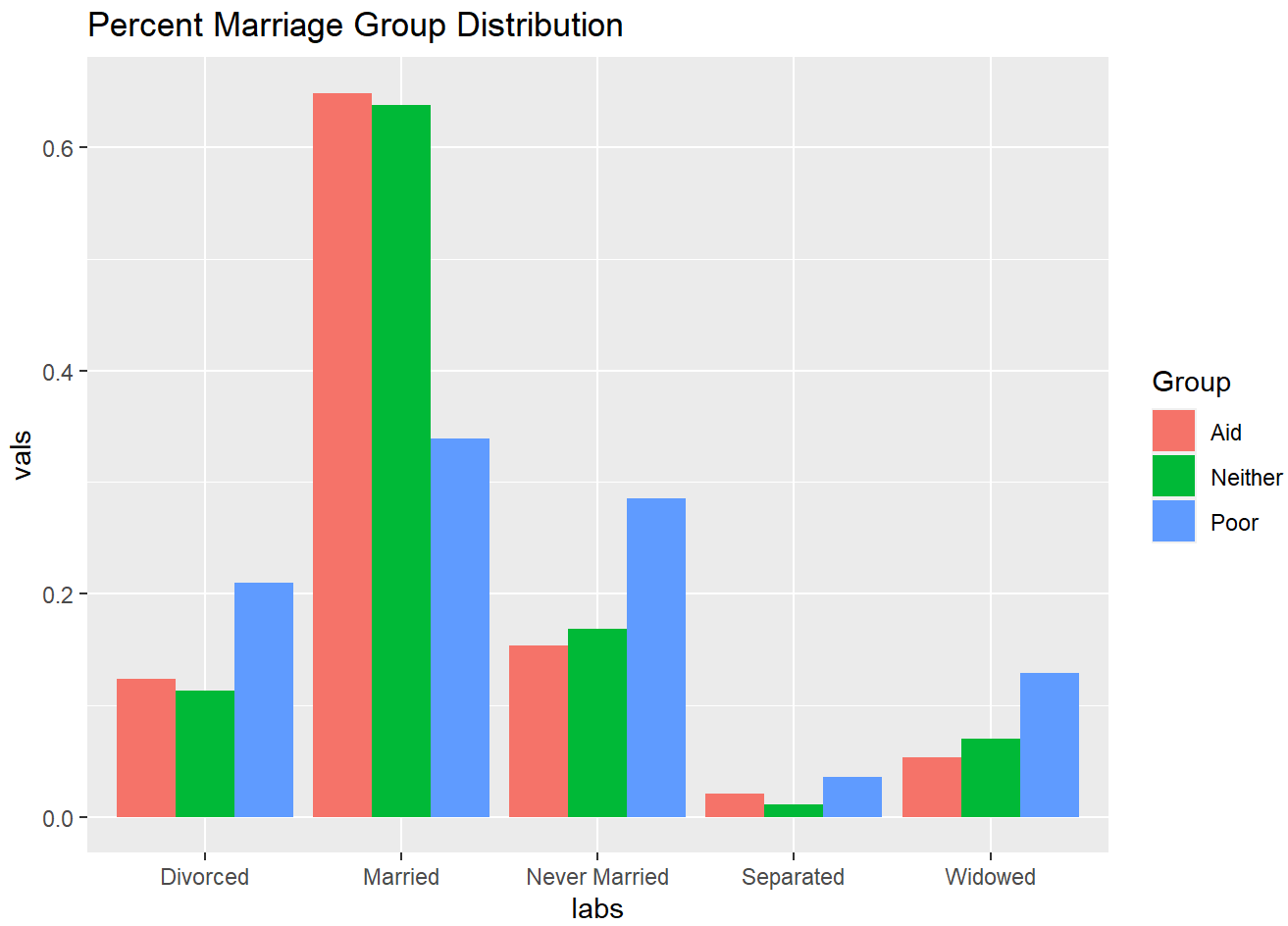
Next, we look at our other demographic variables. The code here is exactly the same as the above, except using the new categories.

```
for (i in c("poor", "aid", "no")) {
  y <- length(df$group[df$group == i])
  x1 <- length(df$group[(df$group == i)&(df$mar == "M")])
  x2 <- length(df$group[(df$group == i)&(df$mar == "D")])
  x3 <- length(df$group[(df$group == i)&(df$mar == "S")])
  x4 <- length(df$group[(df$group == i)&(df$mar == "W")])
  x5 <- length(df$group[(df$group == i)&(df$mar == "NM")])

  print(paste("Group", i))
  print(paste("   Percent Married:", x1/y))
  print(paste("   Percent Divorced:", x2/y))
  print(paste("   Percent Seperated:", x3/y))
  print(paste("   Percent Widowed:", x4/y))
  print(paste("   Percent Never Married:", x5/y))
}
```

```
## [1] "Group poor"
## [1] "   Percent Married: 0.339477393857833"
## [1] "   Percent Divorced: 0.210231973496859"
## [1] "   Percent Seperated: 0.0352878097625583"
## [1] "   Percent Widowed: 0.129188607626399"
## [1] "   Percent Never Married: 0.28581421525635"
## [1] "Group aid"
## [1] "   Percent Married: 0.649106771686493"
## [1] "   Percent Divorced: 0.123625882009619"
## [1] "   Percent Seperated: 0.0207745441631129"
## [1] "   Percent Widowed: 0.0533872197735338"
## [1] "   Percent Never Married: 0.153105582367241"
## [1] "Group no"
## [1] "   Percent Married: 0.638386958638182"
## [1] "   Percent Divorced: 0.112712326561087"
## [1] "   Percent Seperated: 0.0108365693956951"
## [1] "   Percent Widowed: 0.0697635264966712"
## [1] "   Percent Never Married: 0.168300618908364"
```

We see that the poverty group has the least percentage of married individuals and the greatest percentage of all other categories. Here, the aid and no group shave very similar distributions, though the no group has slightly more widowed/never married individuals.

```
i <- "poor"
yp <- length(df$group[df$group == i])
x1p <- length(df$group[(df$group == i)&(df$mar == "M")])/yp
x2p <- length(df$group[(df$group == i)&(df$mar == "D")])/yp
x3p <- length(df$group[(df$group == i)&(df$mar == "S")])/yp
x4p <- length(df$group[(df$group == i)&(df$mar == "W")])/yp
x5p <- length(df$group[(df$group == i)&(df$mar == "NM")])/yp

i <- "aid"
ya <- length(df$group[df$group == i])
x1a <- length(df$group[(df$group == i)&(df$mar == "M")])/ya
x2a <- length(df$group[(df$group == i)&(df$mar == "D")])/ya
x3a <- length(df$group[(df$group == i)&(df$mar == "S")])/ya
x4a <- length(df$group[(df$group == i)&(df$mar == "W")])/ya
x5a <- length(df$group[(df$group == i)&(df$mar == "NM")])/ya

i <- "no"
yn <- length(df$group[df$group == i])
x1n <- length(df$group[(df$group == i)&(df$mar == "M")])/yn
x2n <- length(df$group[(df$group == i)&(df$mar == "D")])/yn
x3n <- length(df$group[(df$group == i)&(df$mar == "S")])/yn
x4n <- length(df$group[(df$group == i)&(df$mar == "W")])/yn
x5n <- length(df$group[(df$group == i)&(df$mar == "NM")])/yn


bar_vals <- c(x1p, x1a, x1n, x2p, x2a, x2n, x3p, x3a, x3n, x4p, x4a, x4n, x5p, x5a, x5n)
bar_labs <- c("Married", "Married", "Married", "Divorced", "Divorced", "Divorced",
              "Separated", "Separated", "Separated", "Widowed", "Widowed", "Widowed", "Never Mar
ried",
              "Never Married", "Never Married")
bar_grps <- c("Poor", "Aid", "Neither", "Poor", "Aid", "Neither", "Poor", "Aid", "Neither",
              "Poor", "Aid", "Neither", "Poor", "Aid", "Neither")

graph_df <- data.frame("vals" = bar_vals, "labs" = bar_labs, "Group" = bar_grps)

gm <- ggplot(graph_df, aes(x=labs, y=vals, fill=Group)) +
  geom_bar(stat="identity", position=position_dodge(preserve="single")) +
  labs(title = "Percent Marriage Group Distribution", xlab = "Marraige Status", ylab = "Percent
Distribution")

gm
```



Here, we can see the above quantitative results visually. Though the aid group is more racially similar to the poor group, for marital status, it is more similar to the no group.

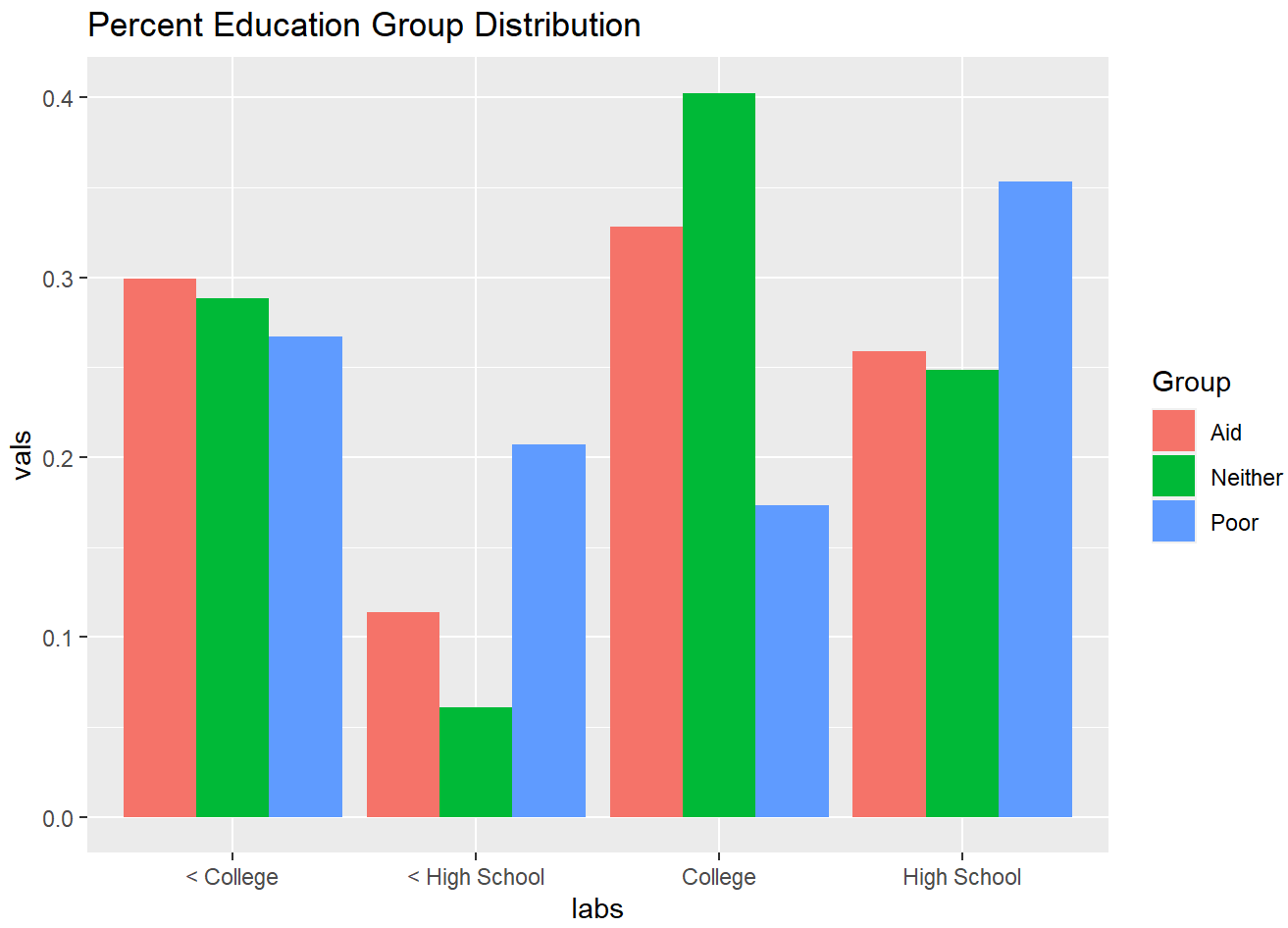
Next, we look at our education statuses.

```
for (i in c("poor", "aid", "no")) {
  y <- length(df$group[df$group == i])
  x1 <- length(df$group[(df$group == i)&(df$edu == "<HS")])
  x2 <- length(df$group[(df$group == i)&(df$edu == "HS")])
  x3 <- length(df$group[(df$group == i)&(df$edu == "<C")])
  x4 <- length(df$group[(df$group == i)&(df$edu == "C")])

  print(paste("Group", i))
  print(paste("   Percent Below High School:", x1/y))
  print(paste("   Percent High School:", x2/y))
  print(paste("   Percent Below College:", x3/y))
  print(paste("   Percent College:", x4/y))
}
```

```
## [1] "Group poor"
## [1] "   Percent Below High School: 0.206997198422026"
## [1] "   Percent High School: 0.353197669257564"
## [1] "   Percent Below College: 0.26675354280662"
## [1] "   Percent College: 0.17305158951379"
## [1] "Group aid"
## [1] "   Percent Below High School: 0.113717641293605"
## [1] "   Percent High School: 0.258946736667649"
## [1] "   Percent Below College: 0.299444158432742"
## [1] "   Percent College: 0.327891463606005"
## [1] "Group no"
## [1] "   Percent Below High School: 0.0607738001559166"
## [1] "   Percent High School: 0.248629873766118"
## [1] "   Percent Below College: 0.288121424623842"
## [1] "   Percent College: 0.402474901454124"
```

We see that the poor group has the most individuals with educations at or below high school. Strangely, the below college category is fairly similar among the groups, but the college category is greater for the aid group, and even greater for the no group.

```
i <- "poor"
yp <- length(df$group[df$group == i])
x1p <- length(df$group[(df$group == i)&(df$edu == "<HS")])/yp
x2p <- length(df$group[(df$group == i)&(df$edu == "HS")])/yp
x3p <- length(df$group[(df$group == i)&(df$edu == "<C")])/yp
x4p <- length(df$group[(df$group == i)&(df$edu == "C")])/yp

i <- "aid"
ya <- length(df$group[df$group == i])
x1a <- length(df$group[(df$group == i)&(df$edu == "<HS")])/ya
x2a <- length(df$group[(df$group == i)&(df$edu == "HS")])/ya
x3a <- length(df$group[(df$group == i)&(df$edu == "<C")])/ya
x4a <- length(df$group[(df$group == i)&(df$edu == "C")])/ya

i <- "no"
yn <- length(df$group[df$group == i])
x1n <- length(df$group[(df$group == i)&(df$edu == "<HS")])/yn
x2n <- length(df$group[(df$group == i)&(df$edu == "HS")])/yn
x3n <- length(df$group[(df$group == i)&(df$edu == "<C")])/yn
x4n <- length(df$group[(df$group == i)&(df$edu == "C")])/yn


bar_vals <- c(x1p, x1a, x1n, x2p, x2a, x2n, x3p, x3a, x3n, x4p, x4a, x4n)
bar_labs <- c("< High School", "< High School", "< High School",
              "High School", "High School", "High School",
              "< College", "< College", "< College", "College", "College", "College")
bar_grps <- c("Poor", "Aid", "Neither", "Poor", "Aid", "Neither", "Poor", "Aid", "Neither",
              "Poor", "Aid", "Neither")

graph_df <- data.frame("vals" = bar_vals, "labs" = bar_labs, "Group" = bar_grps)

ge <- ggplot(graph_df, aes(x=labs, y=vals, fill=Group)) +
  geom_bar(stat="identity", position=position_dodge(preserve="single")) +
  labs(title = "Percent Education Group Distribution",
       xlab = "Education Level", ylab = "Percent Distribution") +
  theme(axis.text.x = element_text(angle = 0))
ge
```



Here, we see the above remarks visually. We have a very small percent of no group individuals with less than a high school education, at about half the rate of aid group, and almost a fifth the rate of the poor group.

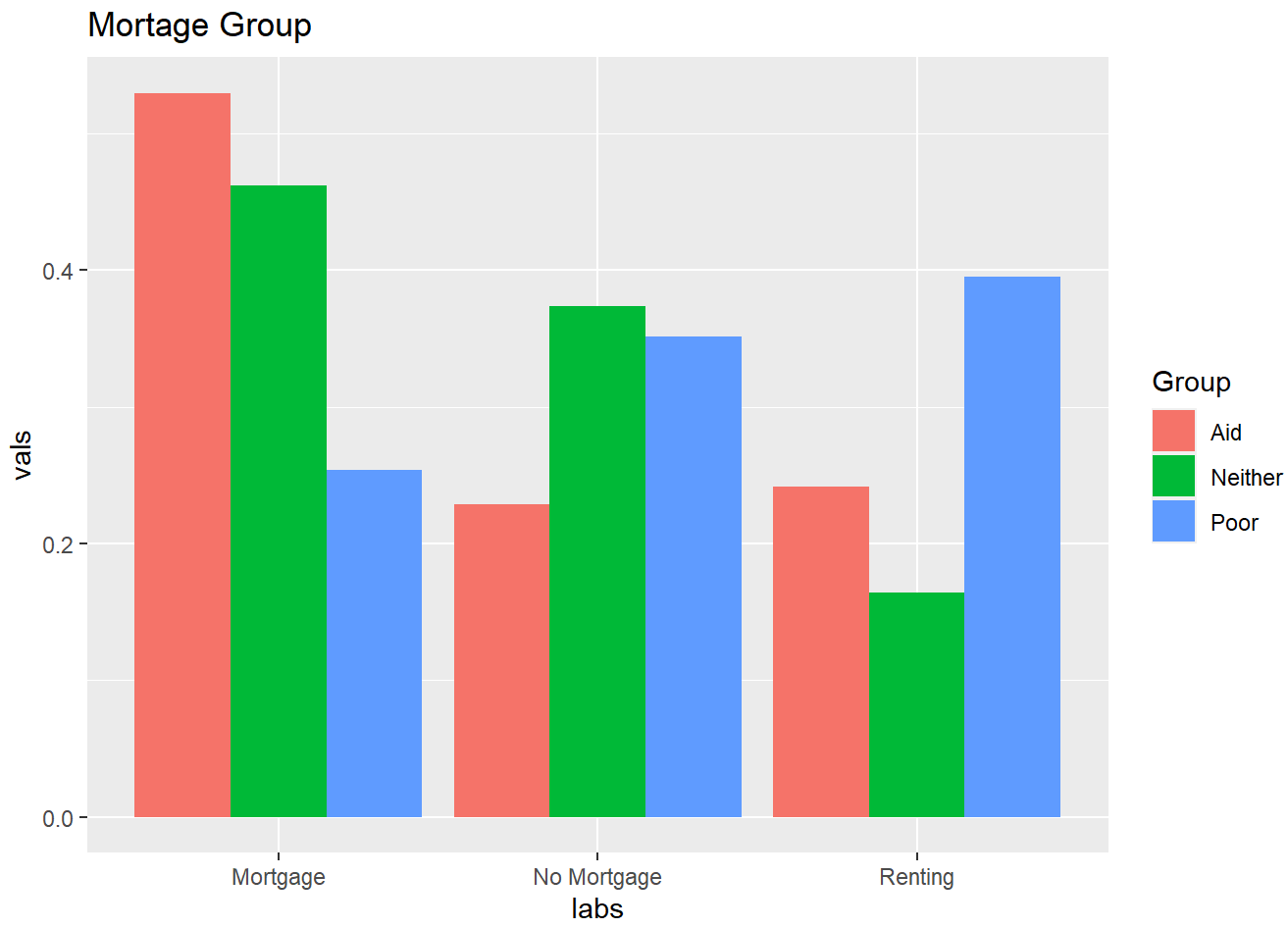
Next, we look at the mortgage categories of each group.

```
for (i in c("poor", "aid", "no")) {
  y <- length(df$group[df$group == i])
  x1 <- length(df$group[(df$group == i)&(df$mortgage == "M")])
  x2 <- length(df$group[(df$group == i)&(df$mortgage == "N")])
  x3 <- length(df$group[(df$group == i)&(df$mortgage == "R")])

  print(paste("Group", i))
  print(paste("   Percent w/ Mortgage:", x1/y))
  print(paste("   Percent w/o Mortage:", x2/y))
  print(paste("   Percent Renting:", x3/y))
}
```

```
## [1] "Group poor"
## [1] "   Percent w/ Mortgage: 0.253640453507654"
## [1] "   Percent w/o Mortage: 0.351351255383894"
## [1] "   Percent Renting: 0.395008291108452"
## [1] "Group aid"
## [1] "   Percent w/ Mortgage: 0.529830714720304"
## [1] "   Percent w/o Mortage: 0.228880635845311"
## [1] "   Percent Renting: 0.241288649434385"
## [1] "Group no"
## [1] "   Percent w/ Mortgage: 0.462409826405537"
## [1] "   Percent w/o Mortage: 0.373614227206348"
## [1] "   Percent Renting: 0.163975946388115"
```

We see that the poor group is most frequently renting, the aid group is most frequently with mortgage, and the no group has the highest percentage without a mortgage. This seems to make sense with what I would assume about these groups. What is interesting is that the poor group has a greater percent without a mortgage than the aid group. Potentially, this could be a result of a failure to explain the questions when giving the survey.

```
i <- "poor"
yp <- length(df$group[df$group == i])
x1p <- length(df$group[(df$group == i)&(df$mortgage == "M")])/yp
x2p <- length(df$group[(df$group == i)&(df$mortgage == "N")])/yp
x3p <- length(df$group[(df$group == i)&(df$mortgage == "R")])/yp

i <- "aid"
ya <- length(df$group[df$group == i])
x1a <- length(df$group[(df$group == i)&(df$mortgage == "M")])/ya
x2a <- length(df$group[(df$group == i)&(df$mortgage == "N")])/ya
x3a <- length(df$group[(df$group == i)&(df$mortgage == "R")])/ya

i <- "no"
yn <- length(df$group[df$group == i])
x1n <- length(df$group[(df$group == i)&(df$mortgage == "M")])/yn
x2n <- length(df$group[(df$group == i)&(df$mortgage == "N")])/yn
x3n <- length(df$group[(df$group == i)&(df$mortgage == "R")])/yn


bar_vals <- c(x1p, x1a, x1n, x2p, x2a, x2n, x3p, x3a, x3n)
bar_labs <- c("Mortgage", "Mortgage", "Mortgage", "No Mortgage", "No Mortgage", "No Mortgage",
              "Renting", "Renting", "Renting")
bar_grps <- c("Poor", "Aid", "Neither", "Poor", "Aid", "Neither", "Poor", "Aid", "Neither")

graph_df <- data.frame("vals" = bar_vals, "labs" = bar_labs, "Group" = bar_grps)

g1 <- ggplot(graph_df, aes(x=labs, y=vals, fill=Group)) +
  geom_bar(stat="identity", position=position_dodge(preserve="single")) +
  labs(title = "Mortage Group", xlab = "Mortgage Status", ylab = "Percent Distribution") +
  theme(axis.text.x = element_text(angle = 0))
g1
```



Here, we see how visually similar the poor and no groups are for the no-mortgage class. However, the poor group seems to have almost double the percent of individuals renting than the other groups.

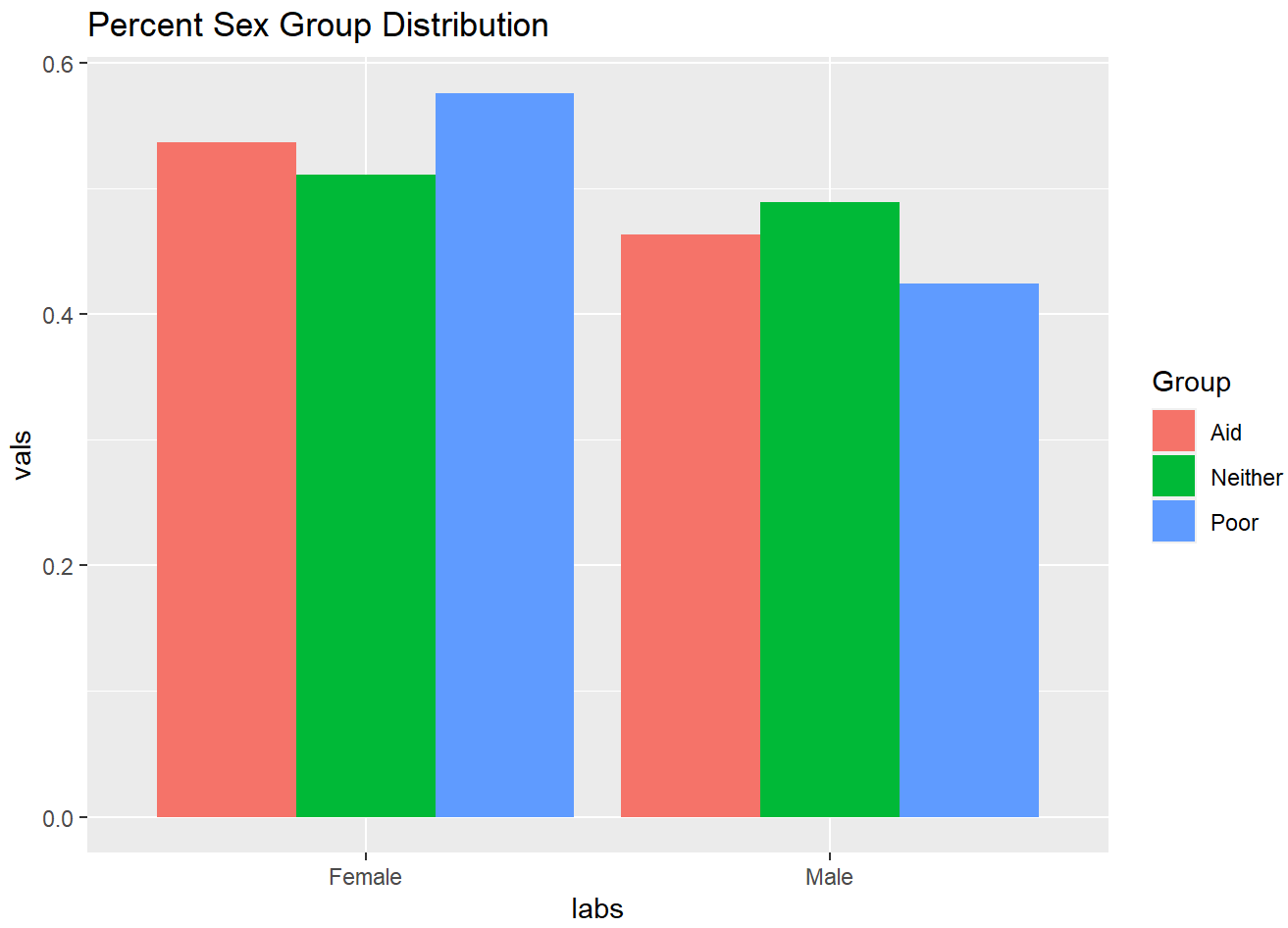
Next, we look at the biological sexes of our three groups.

```
for (i in c("poor", "aid", "no")) {
  y <- length(df$group[df$group == i])
  x1 <- length(df$group[(df$group == i)&(df$sex == "M")])
  x2 <- length(df$group[(df$group == i)&(df$sex == "F")])

  print(paste("Group", i))
  print(paste("   Percent Male:", x1/y))
  print(paste("   Percent Female:", x2/y))
}
```

```
## [1] "Group poor"
## [1] "   Percent Male: 0.424096411210573"
## [1] "   Percent Female: 0.575903588789427"
## [1] "Group aid"
## [1] "   Percent Male: 0.46317652856431"
## [1] "   Percent Female: 0.53682347143569"
## [1] "Group no"
## [1] "   Percent Male: 0.489323373215286"
## [1] "   Percent Female: 0.510676626784714"
```

Quantitatively, we see that the poor group has the greatest percent of females. The aid group has slightly greater percent of females than the no group. However, all groups have a higher percent of females than male.s

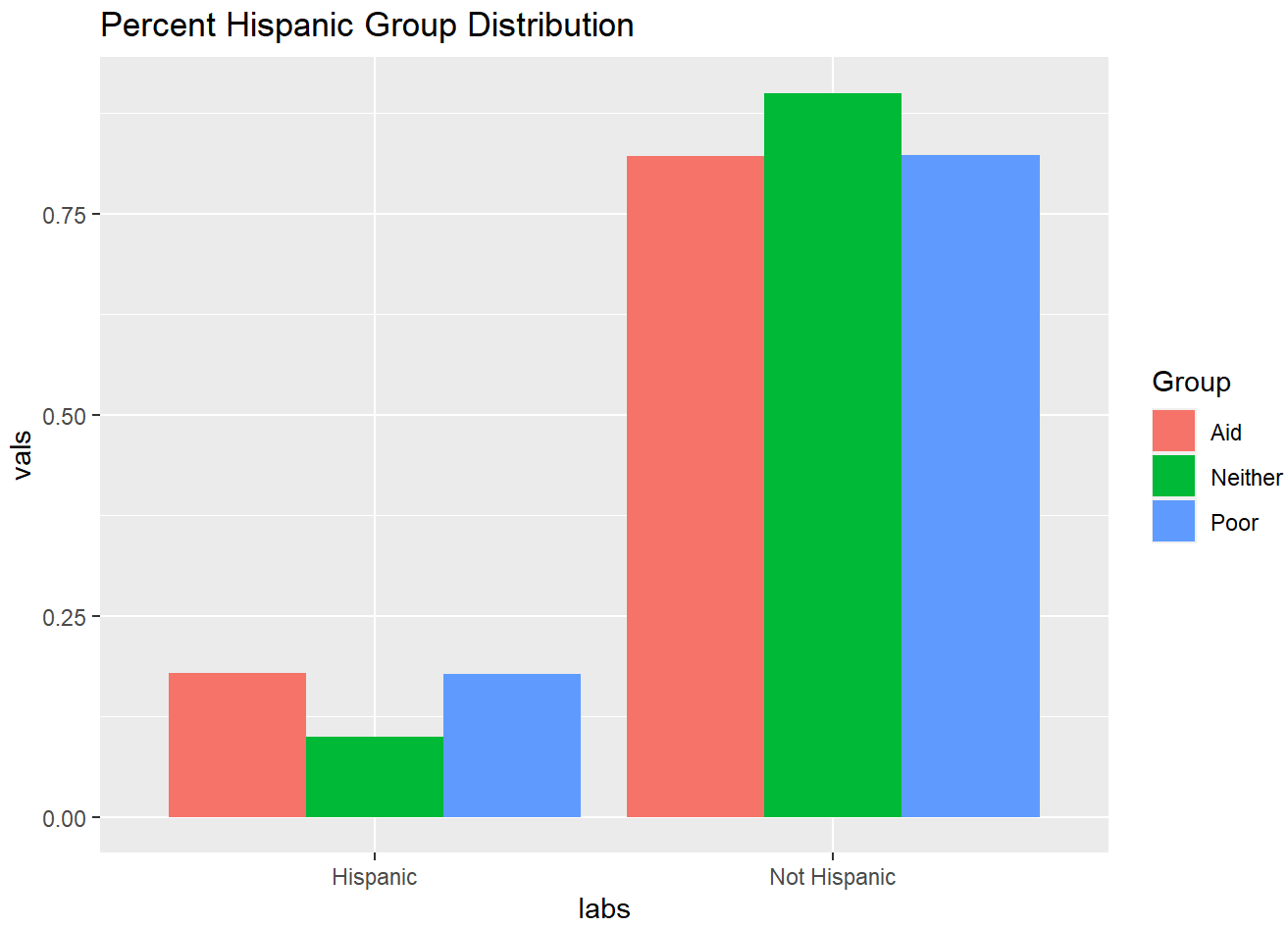
```
i <- "poor"
yp <- length(df$group[df$group == i])
x1p <- length(df$group[(df$group == i)&(df$sex == "M")])/yp
x2p <- length(df$group[(df$group == i)&(df$sex == "F")])/yp

i <- "aid"
ya <- length(df$group[df$group == i])
x1a <- length(df$group[(df$group == i)&(df$sex == "M")])/ya
x2a <- length(df$group[(df$group == i)&(df$sex == "F")])/ya

i <- "no"
yn <- length(df$group[df$group == i])
x1n <- length(df$group[(df$group == i)&(df$sex == "M")])/yn
x2n <- length(df$group[(df$group == i)&(df$sex == "F")])/yn


bar_vals <- c(x1p, x1a, x1n, x2p, x2a, x2n)
bar_labs <- c("Male", "Male", "Male", "Female", "Female", "Female")
bar_grps <- c("Poor", "Aid", "Neither", "Poor", "Aid", "Neither")

graph_df <- data.frame("vals" = bar_vals, "labs" = bar_labs, "Group" = bar_grps)

gs <- ggplot(graph_df, aes(x=labs, y=vals, fill=Group)) +
  geom_bar(stat="identity", position=position_dodge(preserve="single")) +
  labs(title = "Percent Sex Group Distribution", xlab = "Sex", ylab = "Percent Distribution") +
  theme(axis.text.x = element_text(angle = 0))
gs
```



Here, we see the above result visually. Next, we can look at the distribution of hispanic individuals.

```
## [1] "Group poor"
## [1] "   Percent Hispanic: 0.177223774709101"
## [1] "   Percent Not: 0.822776225290899"
## [1] "Group aid"
## [1] "   Percent Hispanic: 0.178524528333053"
## [1] "   Percent Not: 0.821475471666947"
## [1] "Group no"
## [1] "   Percent Hispanic: 0.0996651087207153"
## [1] "   Percent Not: 0.900334891279285"
```

Here, we see that the main difference falls in comparing the no group to the other two. The poor and aid groups have a much greater percent of hispanic individuals than the no group.



Above, we see the distributional differences mentioned above. If we had the aid and poor groups next to each other, they would appear almost level.

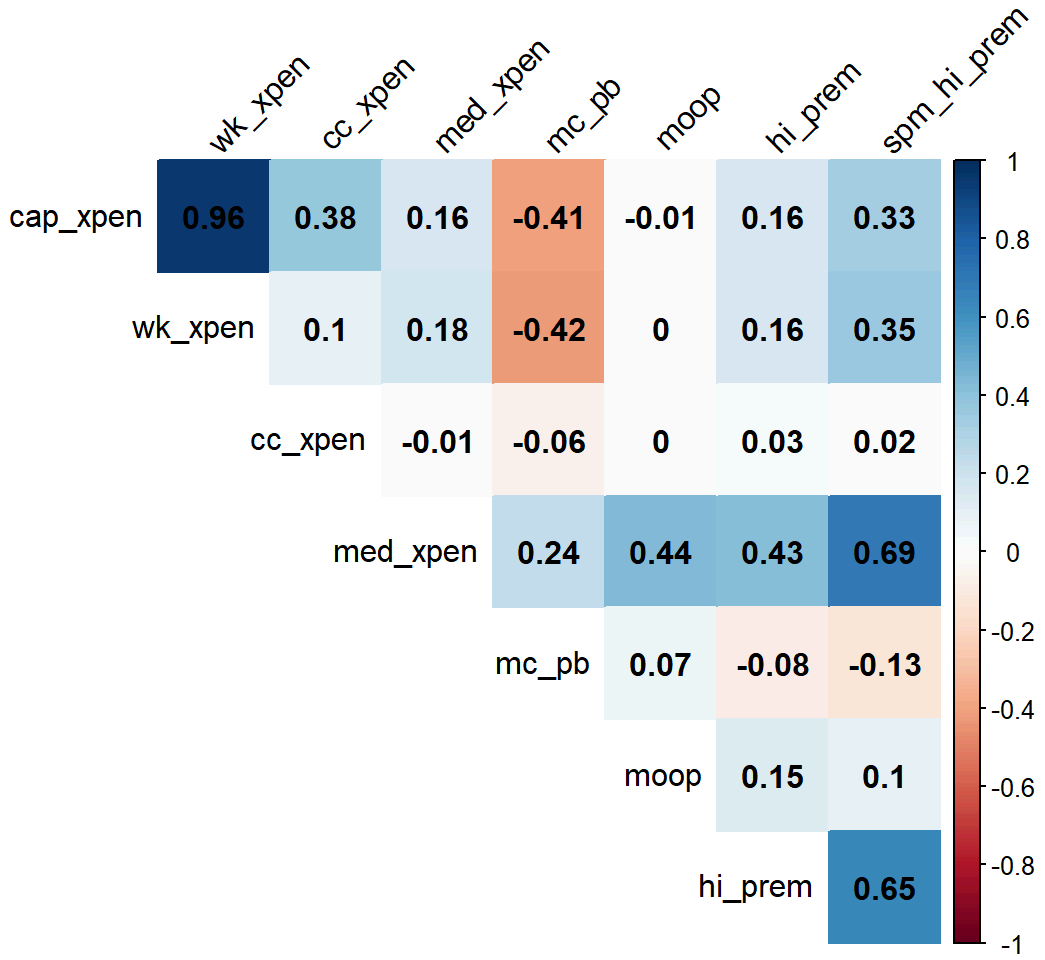
Lastly, we can look at the relative ages of individuals.

Interestingly, it seems that the poor and no group have the most similar distribution of these variables, with peaks around 30 and 65. The aid group has a peak around 40, then a long tail trailing off to the right.We have an abnormally jump around the age of 35 of participants in the survey. Could this represent some bias in how the survey was delivered? We also see a slight peak for all distributions around 95. It could be interesting to see why that is occurring as well.

Expense Variables

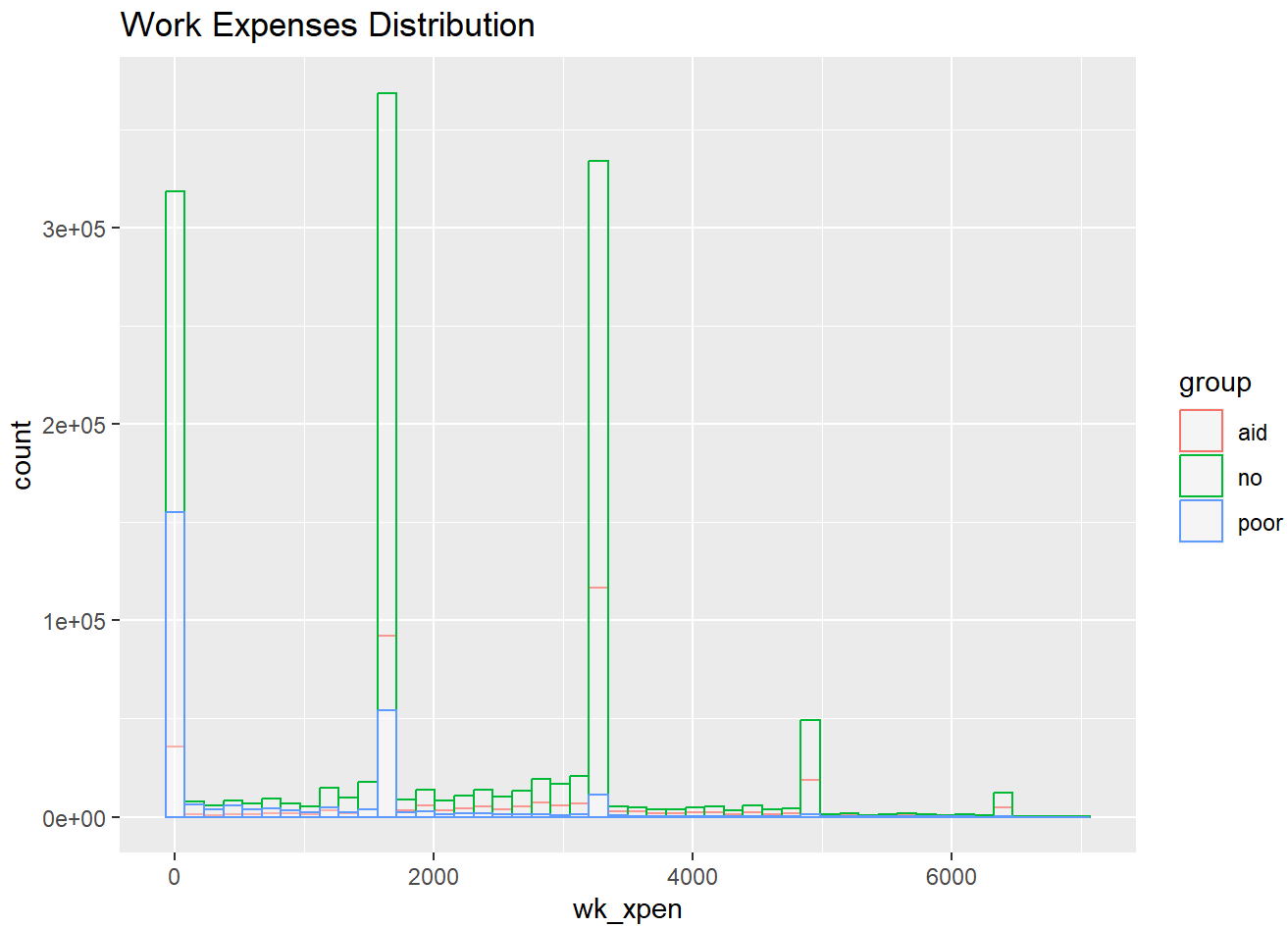
Next we look at our various expense variables. We'll start by examining the correlations between them.

```
# Correlation plot of our expense variables.
corrplot(cor(select(df, cap_xpen, wk_xpen, cc_xpen, med_xpen, mc_pb, moop,
                    hi_prem, spm_hi_prem)),
         method="color", type="upper", diag=FALSE,
         addCoef.col = "black", tl.col="black", tl.srt=45)
```



We see the strongest correlations between the expenses cap and work expenses, which might suggest taht we remove those variables. Interestingly, medical expenses have a greater correlation to spm_hi_prem than hi_prem has to spm_hi_prem. Unlike resource and income, these variables do not seem as tightly correlated. However, though it is not explained in the data dictionary, we believe that spm_hi_prem is a value calculated from hi_prem and our other medical expense variables.

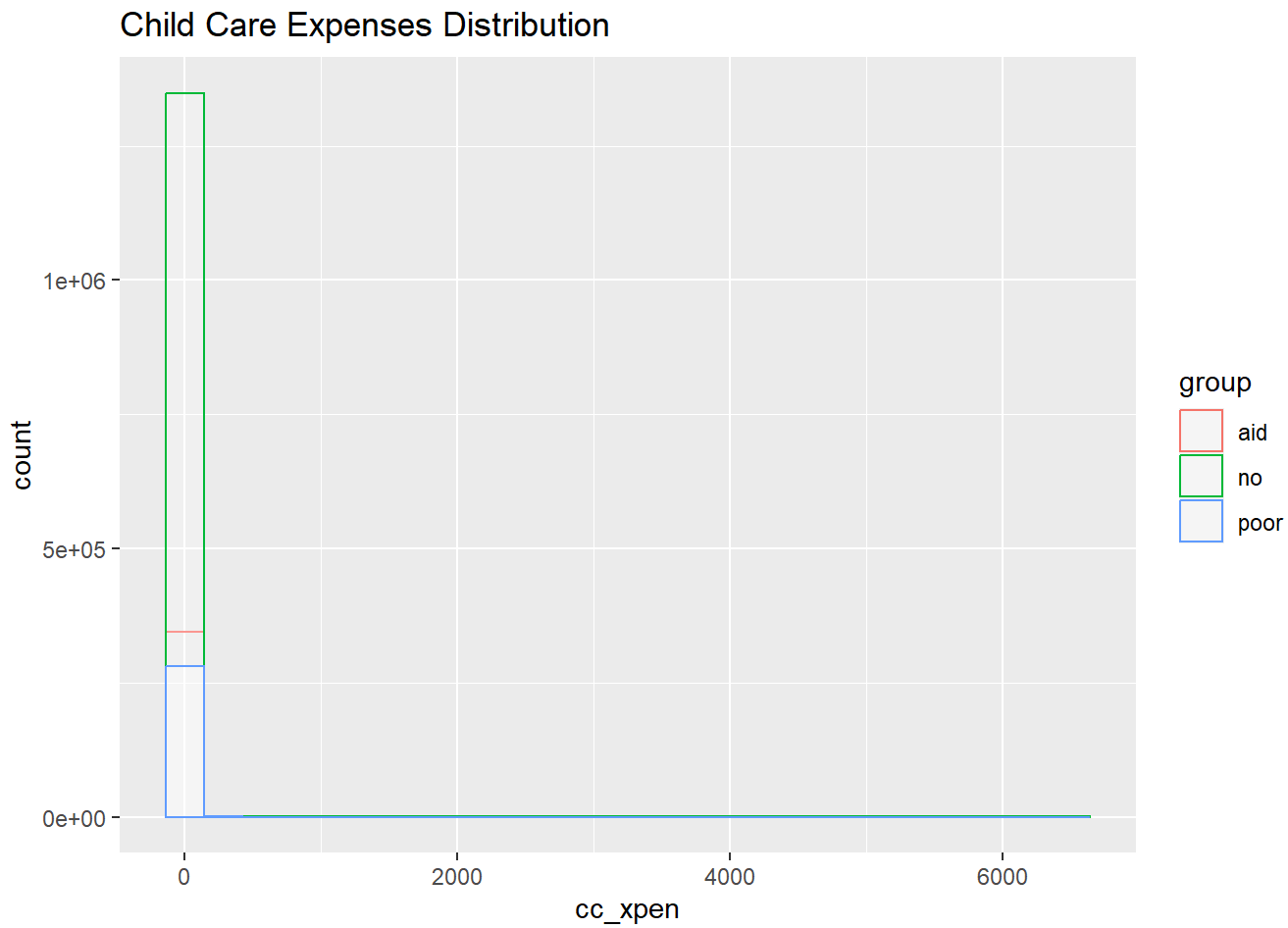
```
# Create a histogram of work expenses by group.
ggplot(df, aes(x=wk_xpen, color=group)) +
  geom_histogram(bins=48, alpha=0.25, position="identity", fill="white") +
  labs(title="Work Expenses Distribution", xlab = "Work Expenses", ylab="Count")
```



Like we saw earlier with tax credit, work expense seems almost discreet in its distribution. We see peaks around 0, 1,600, 3,200, and 4,800, with a final smaller peak around 6,400. Was this meant to be a discreet variable in the data set?

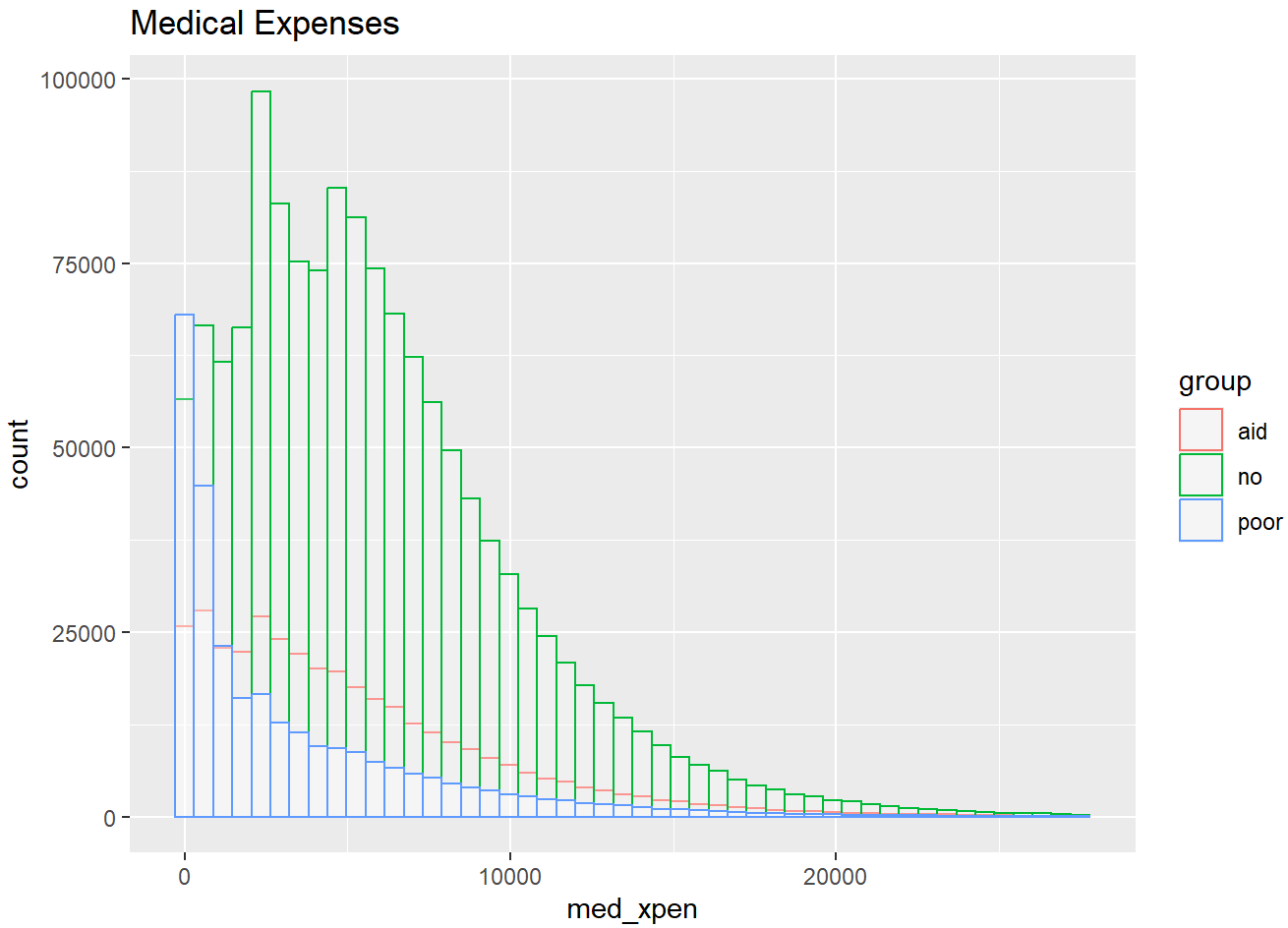
Interestingly, we see that the aid group has the fewest individuals who do not pay any work expenses. I would have expected the no group to at least have a comparable percent ratio, but does that not seem to be the case. The most no group observations are in the bar close to 3,200.

```
# Create a histogram of childcare expenses by group.
ggplot(df, aes(x=cc_xpen, color=group)) +
  geom_histogram(bins=24, alpha=0.25, position="identity", fill="white") +
  labs(title="Child Care Expenses Distribution", xlab = "Child Care Expenses", ylab="Count")
```



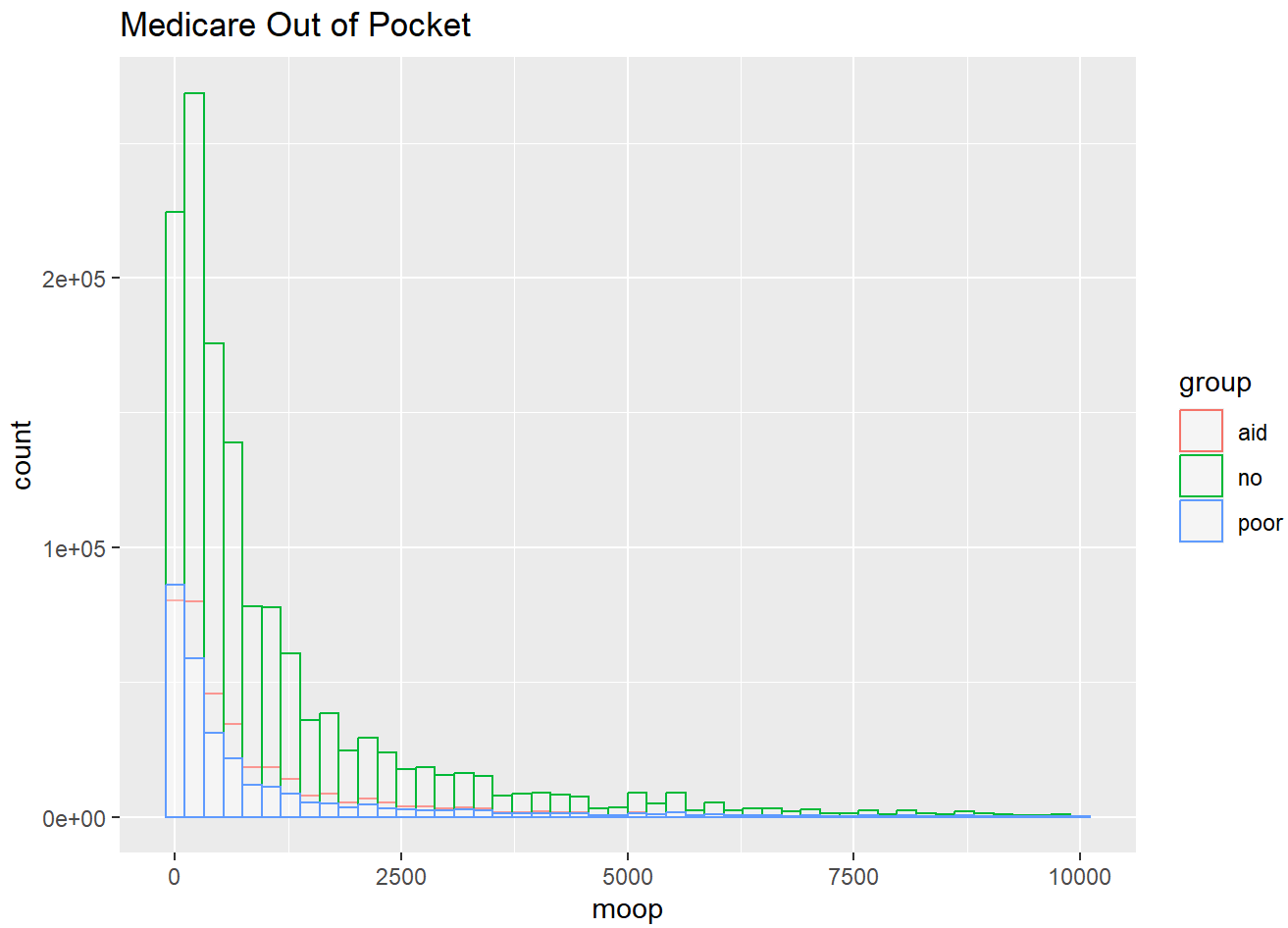
We see that most individuals in our data set do not pay child care expenses, or at least did not report doing so. This might suggest that we should remove this variable from our distribution because there is not enough data.

```
# Create a distribution of medical expenses by group.
g1 <- ggplot(df, aes(x=med_xpen, color=group)) +
  geom_histogram(bins=48, alpha=0.25, position="identity", fill="white") +
  labs(title="Medical Expenses", xlab = "Medical Expenses", ylab="Count")
g1
```



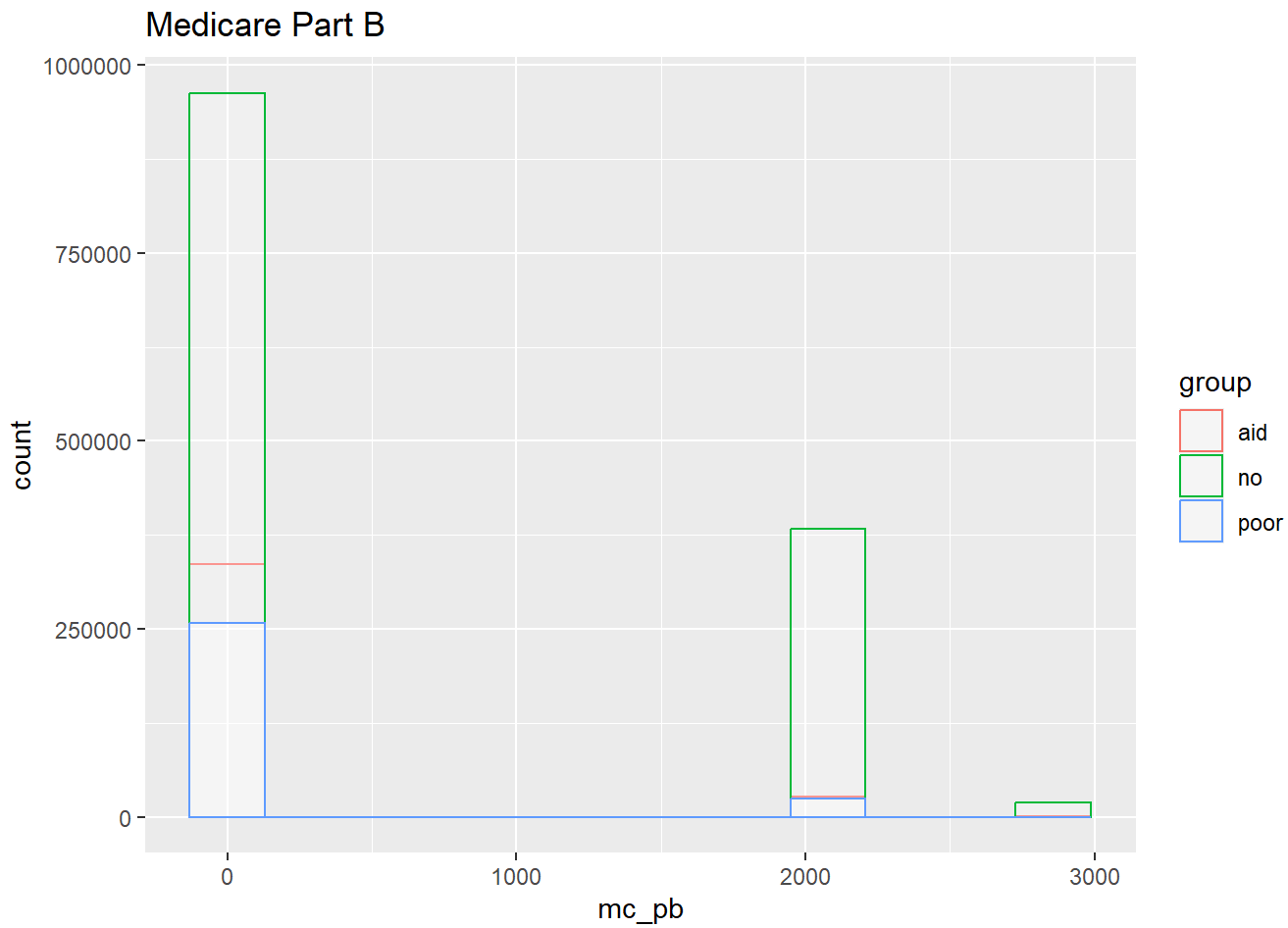
Comparing our groups, we see that the poor group has a long right tail, with most of the distribution centralized at 0. The other two are much more spread out, resembling the results we saw in our income and federal tax distributions.

```
# Create a histogram of MOOP by Group.
g2 <- ggplot(df, aes(x=moop, color=group)) +
  geom_histogram(bins=48, alpha=0.25, position="identity", fill="white") +
  labs(title="Medicare Out of Pocket", xlab = "Medicare Out of Pocket", ylab="Count")
g2
```



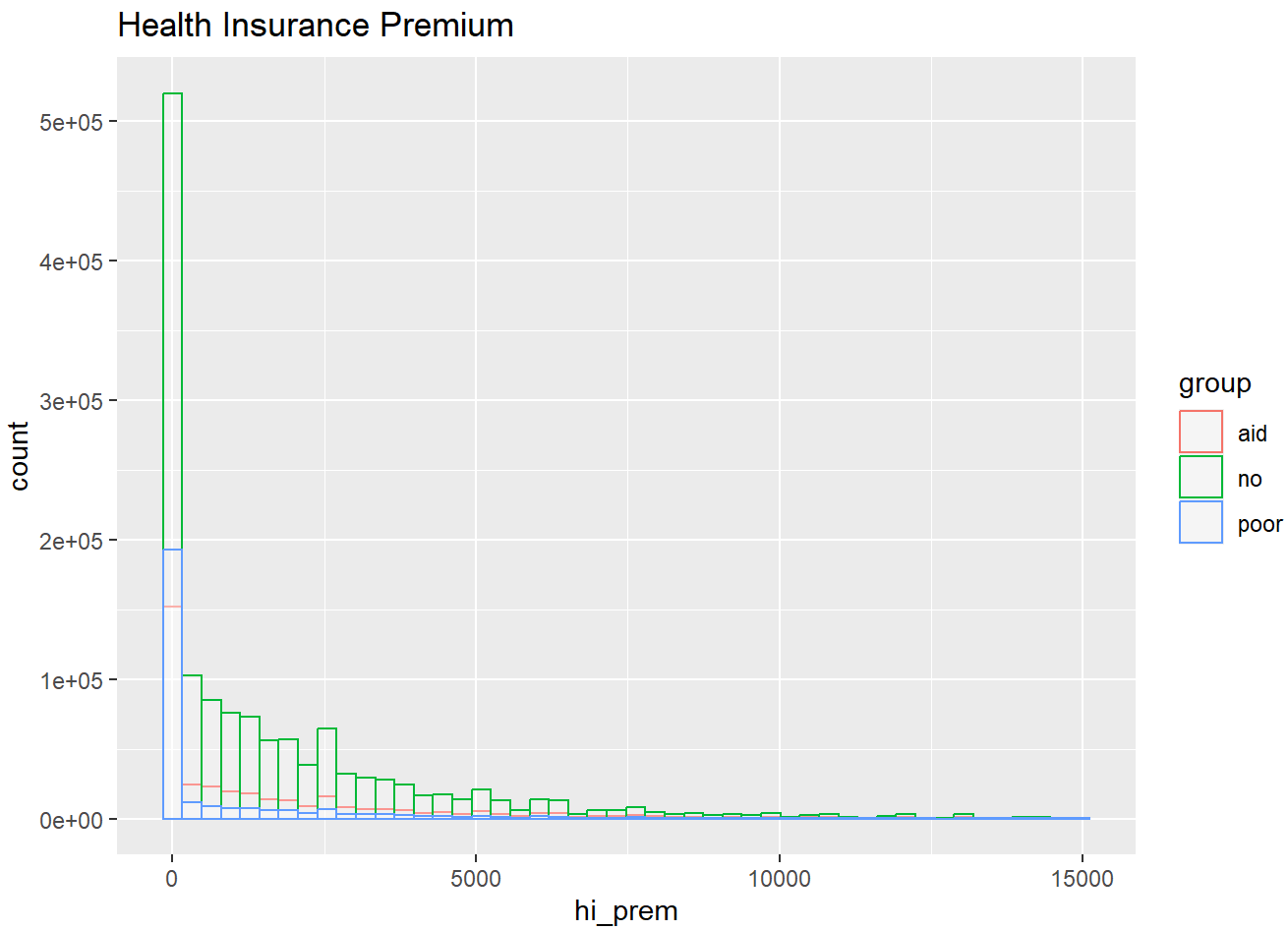
For MOOP, we see that all of our distributions have had their centers pushed toward zero compared to med_xpen. Still, we see that the aid and no group have a greater quantity of observations paying small quantities of moop compared to the poor group.

```
# Create a histogram of Medicare Part B distributions by Group.
g3 <- ggplot(df, aes(x=mc_pb, color=group)) +
  geom_histogram(bins=12, alpha=0.25, position="identity", fill="white") +
  labs(title="Medicare Part B", xlab = "Medicare Part B", ylab="Count")
g3
```



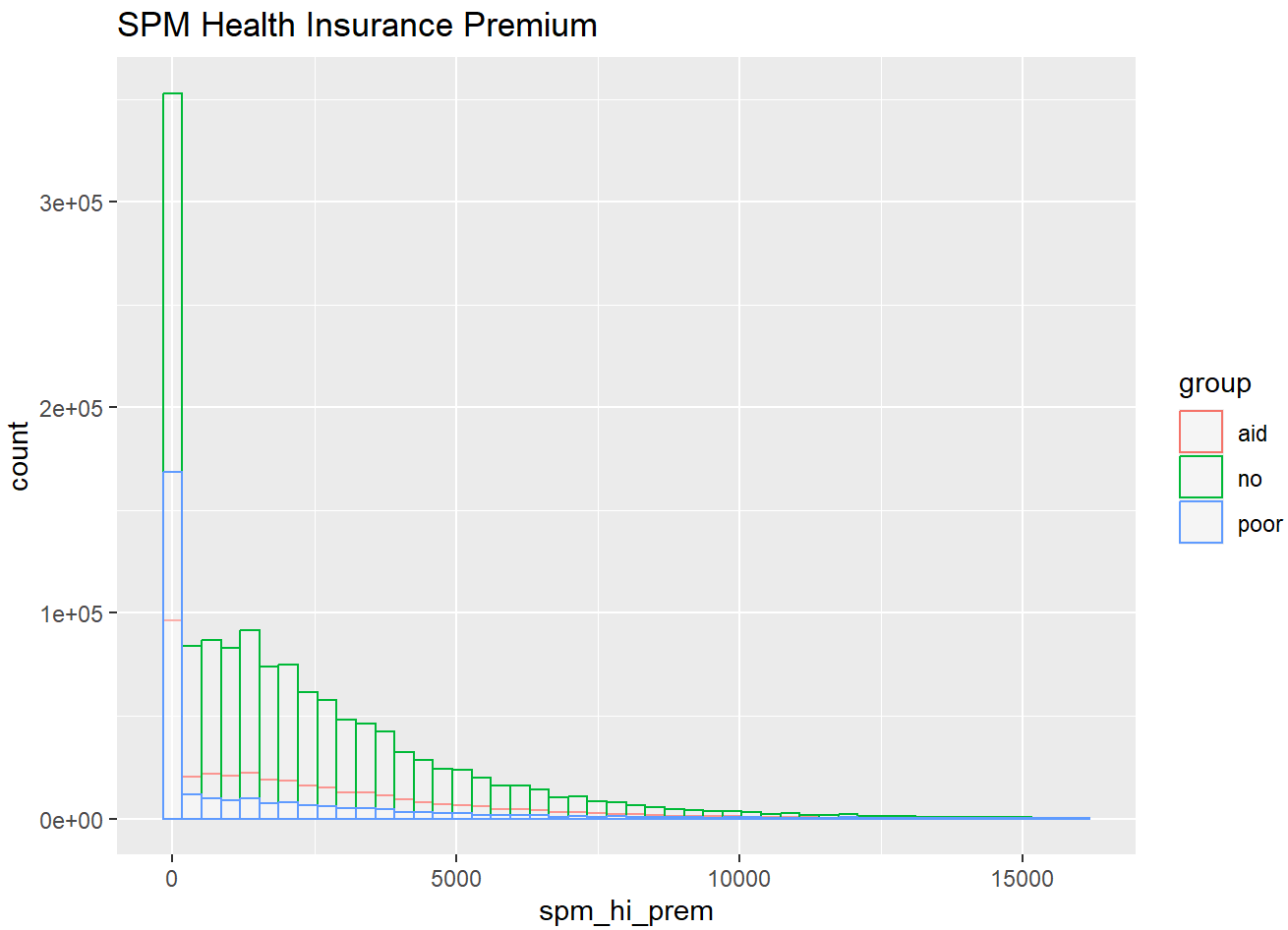
It appears that medicare part b is a mostly discrete value in our data set. We have values around 0, 2,000, and 3,000. There does not seem to be that much difference in overall distribution between our three groups for this variable.

```
# Create a histogram of HI premiums by group.
g4 <- ggplot(df, aes(x=hi_prem, color=group)) +
  geom_histogram(bins=48, alpha=0.25, position="identity", fill="white") +
  labs(title="Health Insurance Premium", xlab = "Health Insurance Premium", ylab="Count")
g4
```



The most interesting result here is that there are more observations of the poor group who have 0 Health Insurance Premiums than there are in the aid group. Because the aid group is the larger group, we would expect it to have greater counts in each bar. However, this is not the case, continuing to show differences between the poor and aid groups.

```
# Create a histogram of SPM HI Premiums by Group.
g5 <- ggplot(df, aes(x=spm_hi_prem, color=group)) +
  geom_histogram(bins=48, alpha=0.25, position="identity", fill="white") +
  labs(title="SPM Health Insurance Premium", xlab = "Health Insurance Premium", ylab="Count")
g5
```

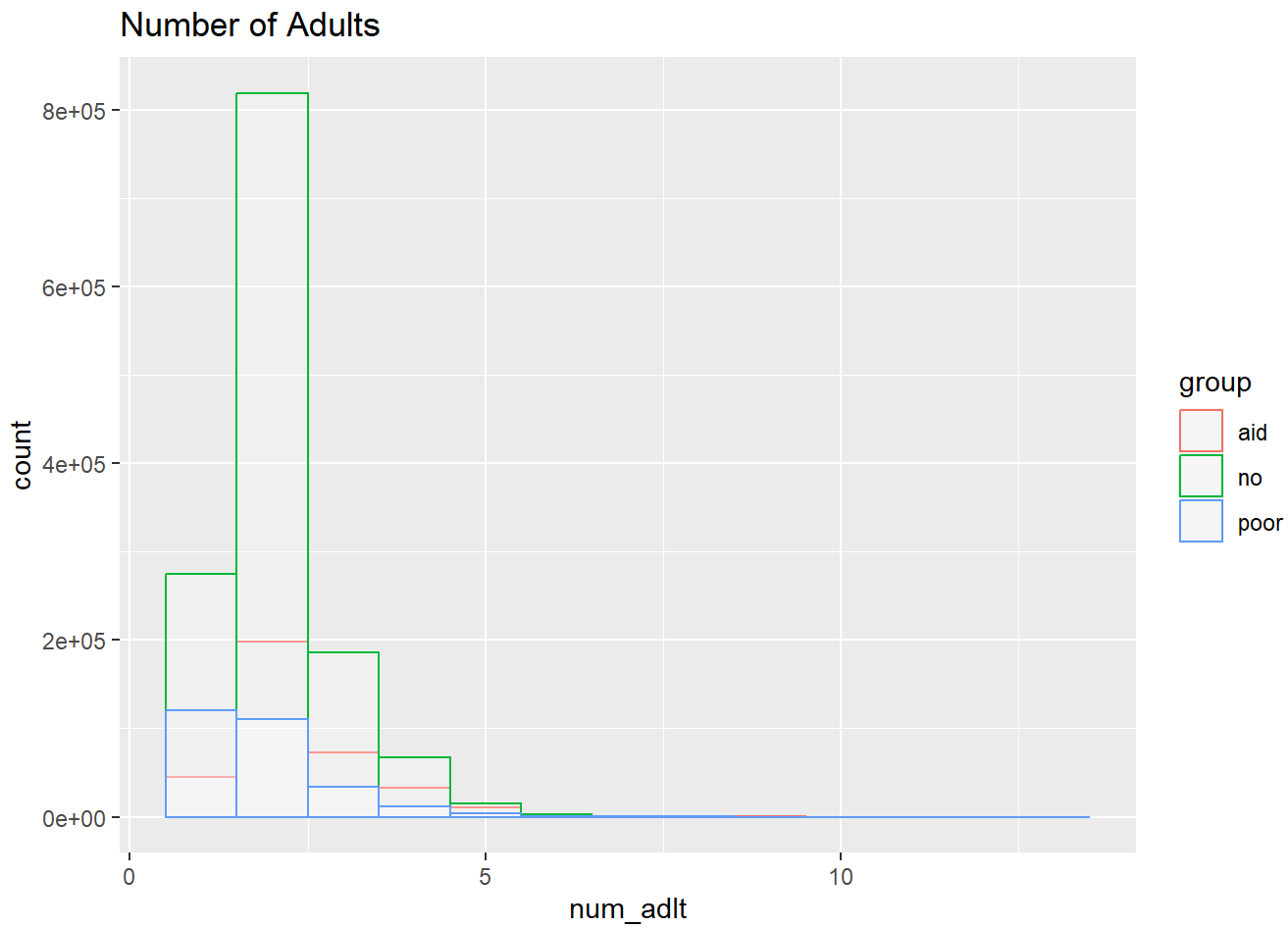


The distribution of SPM_HI_Prem seems slightly smoother than what we saw for HI_Prem. However, the shapes themselves are fairly similar, again with the poor group having more observations at 0 than the aid group.

Household by Group

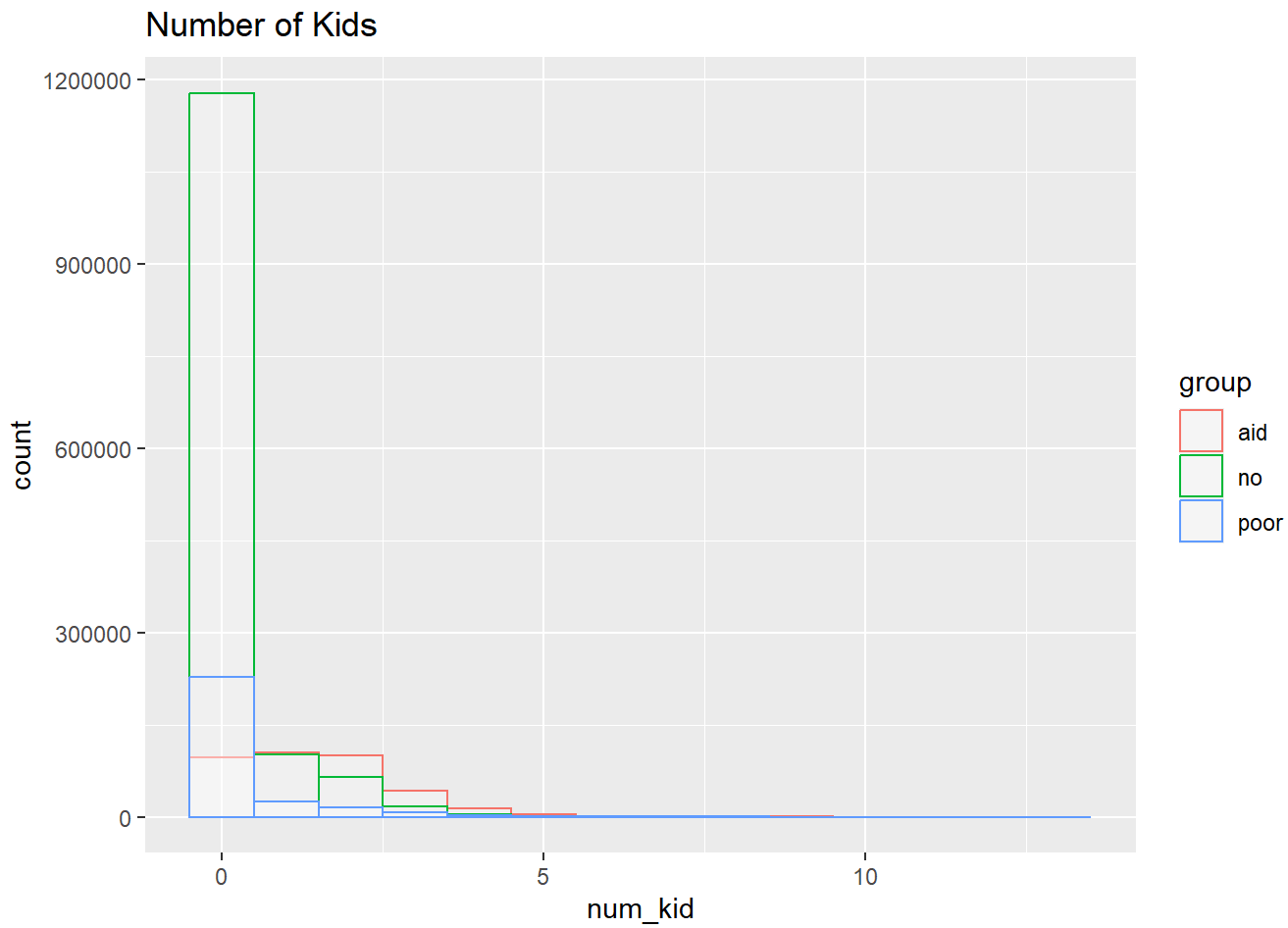
Next, we will examine the household sizes of individuals in each of our groups of interest.

```
# Create a plot of the number of adults by group.
g2 <- ggplot(df, aes(x=num_adlt, color=group)) +
  geom_histogram(bins=13, alpha=0.25, position="identity", fill="white") +
  labs(title="Number of Adults", xlab = "Number of Adults", ylab="Count")
g2
```



Starting with the number of adults in a household, we see that most homes have 1 to 3 adults. Specifically, we see that poor individuals are more likely to have 1 individual in a household than our other two groups. Generally, the distribution for our aid group seems most similar to that of the no group.

```
# Create a plot of the number of kids by group.
g3 <- ggplot(df, aes(x=num_kid, color=group)) +
  geom_histogram(bins=14, alpha=0.25, position="identity", fill="white") +
  labs(title="Number of Kids", xlab = "Number of Kids", ylab="Count")
g3
```



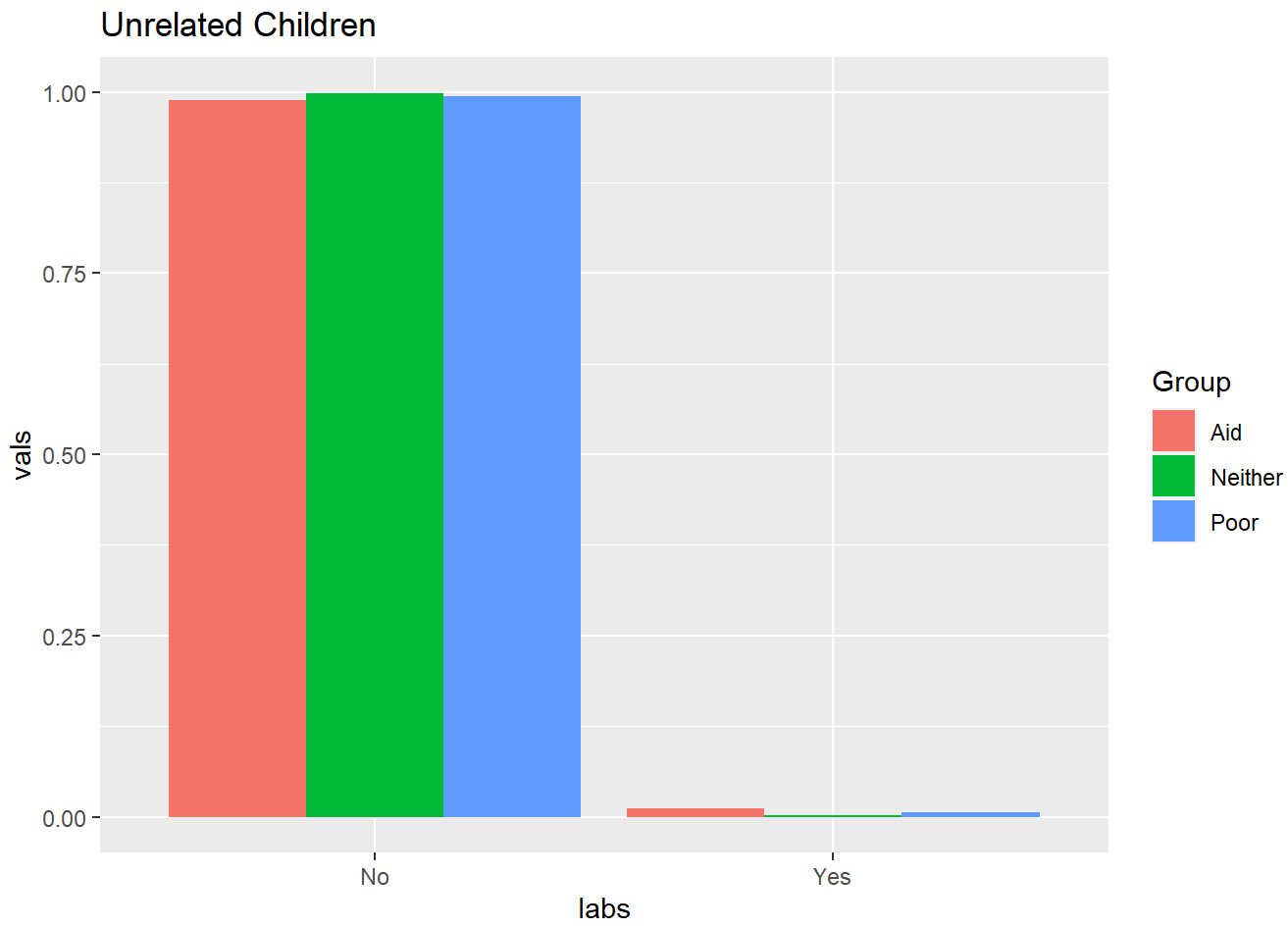
We can look similarly at the number of kids. Interestingly, the aid group seems to have the most children. This could reflect the fact that they are the most likely to earn school lunch subsidies. Ironically, the no and poor groups seem most similar in this distribution.

```
# Loop through each of our target groups.
for (i in c("poor", "aid", "no")) {
  # Find total group size.
  y <- length(df$group[df$group == i])
  # Find number of observations with and without ui_kids.
  x1 <- length(df$group[(df$group == i)&(df$ui_kids == 1)])
  x2 <- length(df$group[(df$group == i)&(df$ui_kids == 0)])

  # Print results as percentages.
  print(paste("Group", i))
  print(paste("   Percent w/ Unrelated Kids:", x1/y))
  print(paste("   Percent w/o Unrelated Kids:", x2/y))
}
```

```
## [1] "Group poor"
## [1] "   Percent w/ Unrelated Kids: 0.00615708010950655"
## [1] "   Percent w/o Unrelated Kids: 0.993842919890493"
## [1] "Group aid"
## [1] "   Percent w/ Unrelated Kids: 0.0112407187672926"
## [1] "   Percent w/o Unrelated Kids: 0.988759281232707"
## [1] "Group no"
## [1] "   Percent w/ Unrelated Kids: 0.00148376966800013"
## [1] "   Percent w/o Unrelated Kids: 0.998516230332"
```

Looking first at the UI_Kids column quantitatively, we see that the aid and no group is the most likely to have unrelated children in their household. Generally, though, it does not seem that common for households to have unrelated kids, suggesting that this variable might not be useful.

```
# using very similar calculations to our demographic variables like race and
# marital status, we find the distribution of unidentified kids.

# Calculate percents for poor individuals.
i <- "poor"
yp <- length(df$group[df$group == i])
x1p <- length(df$group[(df$group == i)&(df$ui_kids == 1)])/yp
x2p <- length(df$group[(df$group == i)&(df$ui_kids == 0)])/yp

# Percents for aid group individuals.
i <- "aid"
ya <- length(df$group[df$group == i])
x1a <- length(df$group[(df$group == i)&(df$ui_kids == 1)])/ya
x2a <- length(df$group[(df$group == i)&(df$ui_kids == 0)])/ya

# Percents for no group individuals.
i <- "no"
yn <- length(df$group[df$group == i])
x1n <- length(df$group[(df$group == i)&(df$ui_kids == 1)])/yn
x2n <- length(df$group[(df$group == i)&(df$ui_kids == 0)])/yn

# Set up the values for a visualization data frame.
bar_vals <- c(x1p, x1a, x1n, x2p, x2a, x2n)
bar_labs <- c("Yes", "Yes", "Yes", "No", "No", "No")
bar_grps <- c("Poor", "Aid", "Neither", "Poor", "Aid", "Neither")

# Assign values to the data frame.
graph_df <- data.frame("vals" = bar_vals, "labs" = bar_labs, "Group" = bar_grps)

# Create a percentage bar plot for our three groups comparing ui_kids.
g4 <- ggplot(graph_df, aes(x=labs, y=vals, fill=Group)) +
  geom_bar(stat="identity", position=position_dodge(preserve="single")) +
  labs(title = "Unrelated Children",
       xlab = "Is an Unrelated Child in Household?",
       ylab = "Percent Distribution") +
  theme(axis.text.x = element_text(angle = 0))

g4
```



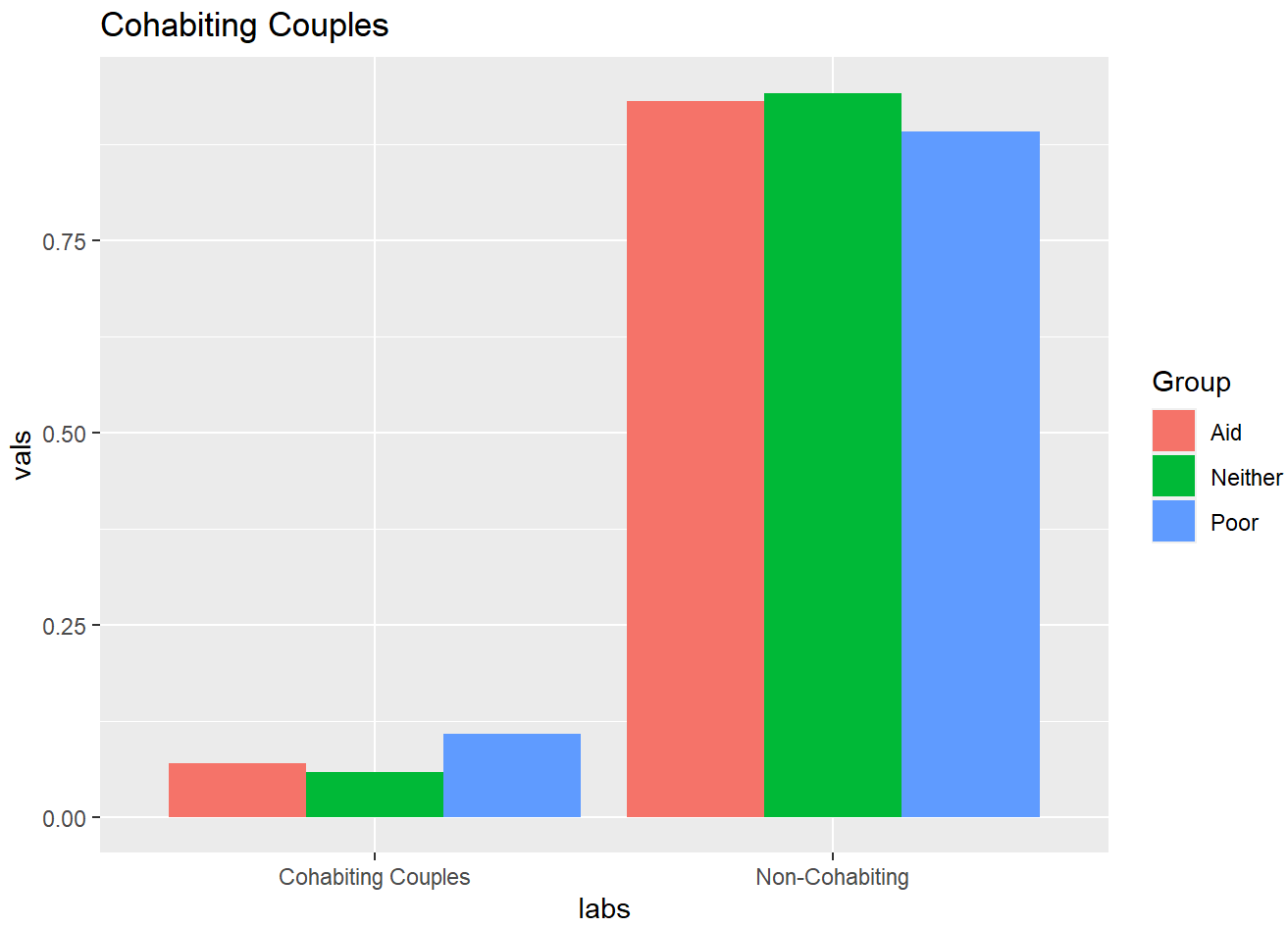
This visualization is really not super helpful because the margins are so small. However, they do show the same trend as we saw in our quantitative metric.

```
# Do the same percentage calculations as done with ui_kids.
for (i in c("poor", "aid", "no")) {
  y <- length(df$group[df$group == i])
  x1 <- length(df$group[(df$group == i)&(df$cohabit == 1)])
  x2 <- length(df$group[(df$group == i)&(df$cohabit == 0)])

  print(paste("Group", i))
  print(paste("   Percent w/ Cohabiting Couples:", x1/y))
  print(paste("   Percent w/o Cohabiting Couples:", x2/y))
}
```

```
## [1] "Group poor"
## [1] "   Percent w/ Cohabiting Couples: 0.108341884833485"
## [1] "   Percent w/o Cohabiting Couples: 0.891658115166515"
## [1] "Group aid"
## [1] "   Percent w/ Cohabiting Couples: 0.0694044869271239"
## [1] "   Percent w/o Cohabiting Couples: 0.930595513072876"
## [1] "Group no"
## [1] "   Percent w/ Cohabiting Couples: 0.0586275679573096"
## [1] "   Percent w/o Cohabiting Couples: 0.94137243204269"
```

Here, we see a more significant number of observations falling into different groups. We see that the poor group has the greatest percent of observations with cohabiting couples. The other two groups have a much closer distribution.

```
# Create visualization of cohabit variable as we did with ui_kids.
i <- "poor"
yp <- length(df$group[df$group == i])
x1p <- length(df$group[(df$group == i)&(df$cohabit == 1)])/yp
x2p <- length(df$group[(df$group == i)&(df$cohabit == 0)])/yp

i <- "aid"
ya <- length(df$group[df$group == i])
x1a <- length(df$group[(df$group == i)&(df$cohabit == 1)])/ya
x2a <- length(df$group[(df$group == i)&(df$cohabit == 0)])/ya

i <- "no"
yn <- length(df$group[df$group == i])
x1n <- length(df$group[(df$group == i)&(df$cohabit == 1)])/yn
x2n <- length(df$group[(df$group == i)&(df$cohabit == 0)])/yn


bar_vals <- c(x1p, x1a, x1n, x2p, x2a, x2n)
bar_labs <- c("Cohabiting Couples", "Cohabiting Couples", "Cohabiting Couples",
              "Non-Cohabiting", "Non-Cohabiting", "Non-Cohabiting")
bar_grps <- c("Poor", "Aid", "Neither", "Poor", "Aid", "Neither")

graph_df <- data.frame("vals" = bar_vals, "labs" = bar_labs, "Group" = bar_grps)

g5 <- ggplot(graph_df, aes(x=labs, y=vals, fill=Group)) +
  geom_bar(stat="identity", position=position_dodge(preserve="single")) +
  labs(title = "Cohabiting Couples",
       xlab = "Cohabitation Status", ylab = "Percent Distribution") +
  theme(axis.text.x = element_text(angle = 0))
g5
```



Again, this visualization does not tell us a lot more than the raw percents.

State by Group

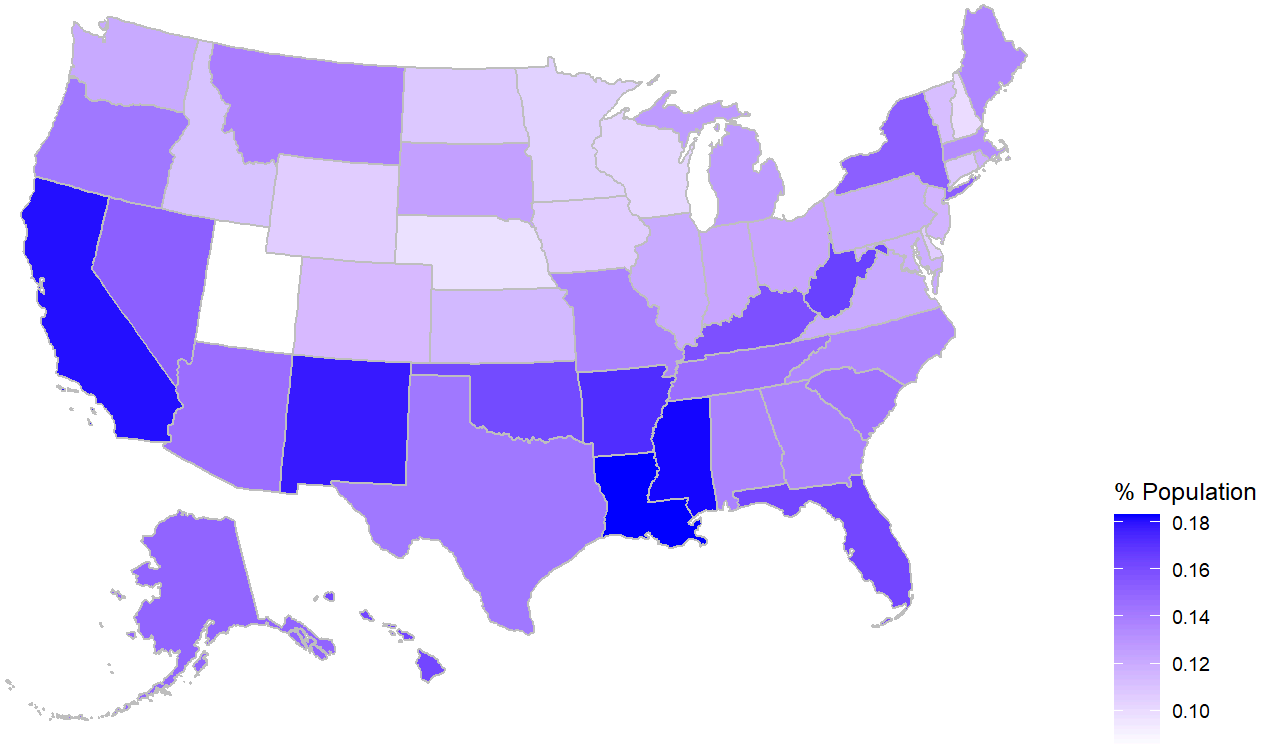
Since our data is distributed across the United States, we can use another R Package to create spatial visualizations of distributions by state.

```
# Store all states in our data set.
states <- unique(df$st)
# Create an empty data frame to store with counts.
state_df <- data.frame("poor" = rep(0, 51),
                       "aid" = rep(0, 51),
                       "no" = rep(0, 51), "state" = states,
                       row.names = states)

# For each state.
for (i in states) {
  # Find the number of individuals in each state.
  st_size <- length(df$st[df$st == i])
  # Find the percent of poor individuals out of total population.
  state_df[i, "poor"] <- length(df$st[(df$st == i)&
                                        (df$group == "poor")])/st_size
  # Find the percent of aid group individuals in this state.
  state_df[i, "aid"] <- length(df$st[(df$st == i)&
                                       (df$group == "aid")])/st_size
  # Find the percent of no group individuals in this state.
  state_df[i, "no"] <- length(df$st[(df$st == i)&
                                      (df$group == "no")])/st_size
}
```

We'll start by looking at the percent of poor individuals in each state.

```
# Create a plot of the percentage of poor individuals in the US.
p1 <- plot_usmap(data = state_df, values = "poor", color = "gray") +
  scale_fill_continuous(low = "white", high = "blue", name = "% Population",
                        label = scales::comma) +
  labs(title = "Poverty Density in US") +
  theme(legend.position = "right")
p1
```

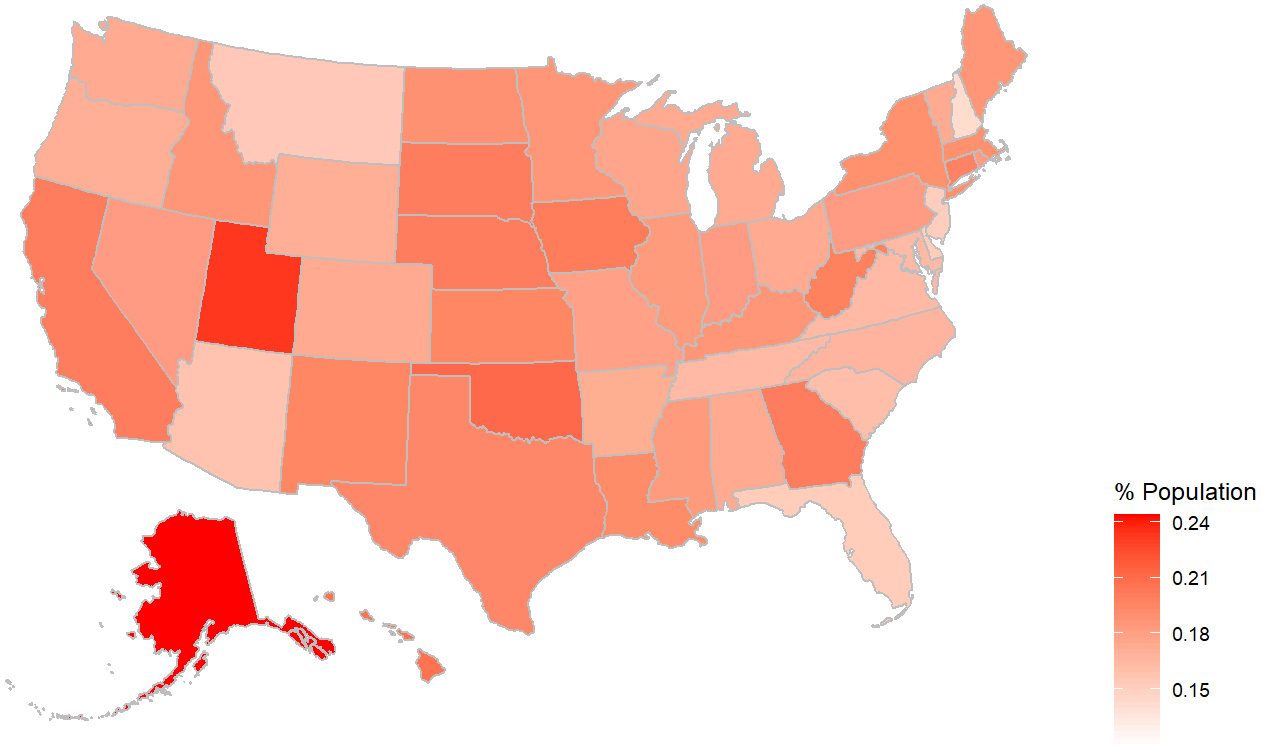
Poverty Density in US



We see that California, New Mexico, Louisiana, and Alabama seem to have the greatest impoverished percentage, whereas Utah and New Hampshire have the least.

```
# Create a visualization of the percentage of aid-group individuals in the US.
p2 <- plot_usmap(data = state_df, values = "aid", color = "gray") +
  scale_fill_continuous(low = "white", high = "red", name = "% Population",
                        label = scales::comma) +
  labs(title = "Non-Impoverished Receiving Aid in US") +
  theme(legend.position = "right")
p2
```

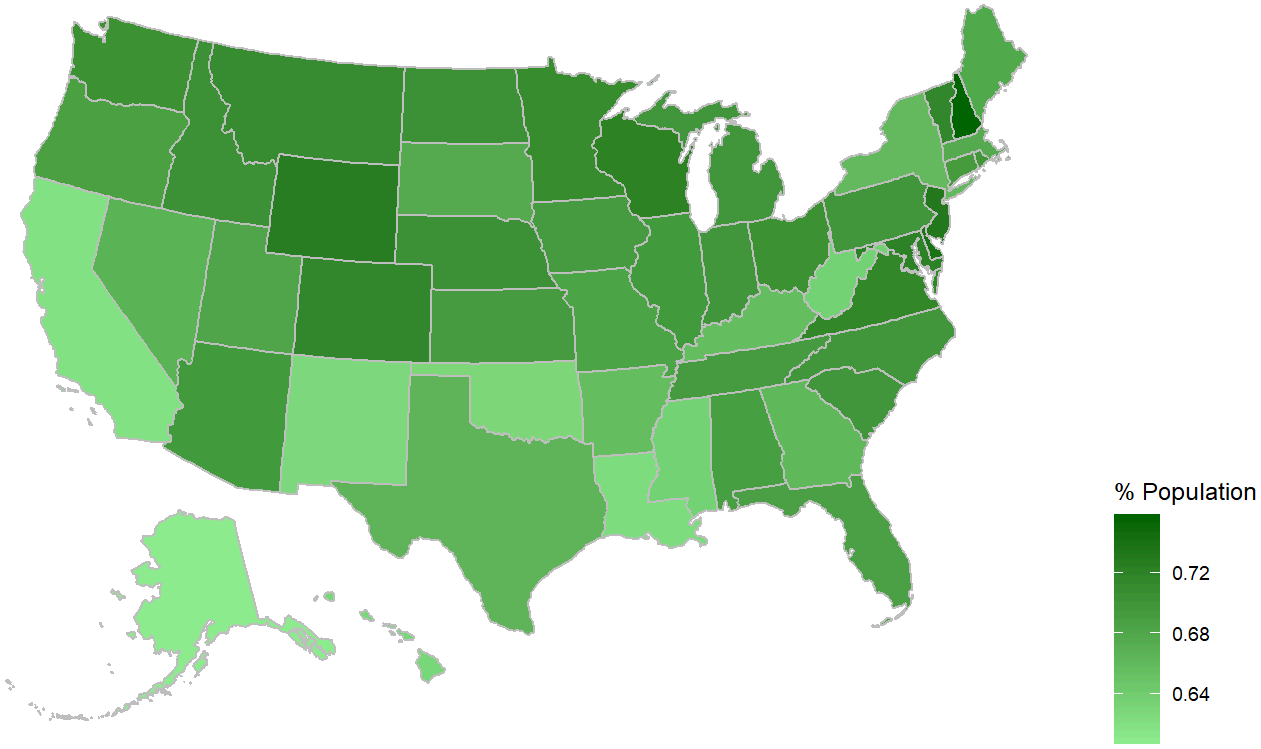
Non-Impoverished Receiving Aid in US



Next, for the aid group, we actually see very high percentages in Utah and Alaska, with lower distributions in New Hampshire, New Jersy, Deleware, and Montana.

```
# Create a visualization of the percent of no-group individuals in the US.
p3 <- plot_usmap(data = state_df, values = "no", color = "gray") +
  scale_fill_continuous(low = "lightgreen", high = "darkgreen",
                        name = "% Population", label = scales::comma) +
  labs(title = "Default Group Distribution in USA") +
  theme(legend.position = "right")
p3
```

Default Group Distribution in USA



Lastly, we see that New Hampshire, New Jersy, and Delaware have the greatest distributions of no-group individuals, whereas states like California, Oklahoma, West Virginia, Louisiana, and Alaska have relatively fewer.

It is very difficult to use state as a variable in our analysis, so these visualizations are likely the most we will see of this variable.