Explore factors affecting income in NLSY ’79 data

Doug Perez

Introduction

Many factors effect a person’s income over a period of years. In this analysis, we sought to investigate whether race plays a significant role in determining an individual’s income. Further, we investigated the data over time to find any trends that may have developed over the past three decades. The income data has been standardized to 2014 dollars, making these comparisons both possible and valid.

To begin, we initialize the environment and load the data.

Examine the income and physical data sets, as those are the ones we will be working with.

```
glimpse(income_data_nlsy79)
```

```
## Observations: 291,778
## Variables: 3
## $ CASEID <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, ...
## $ income <int> NA, 10000, 7000, 1086, 2300, 3250, 4975, 7500, 5000, 90...
## $ year   <int> 1982, 1982, 1982, 1982, 1982, 1982, 1982, 1982, 1982, 1...
```

```
glimpse(physical_data_nlsy79)
```

```
## Observations: 253,720
## Variables: 9
## $ CASEID <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, ...
## $ weight <int> NA, 120, NA, 110, 130, 200, 131, 179, 145, 115, 155, 11...
## $ year   <int> 1981, 1981, 1981, 1981, 1981, 1981, 1981, 1981, 1981, 1...
## $ eyes   <chr> NA, "hazel", "blue", "blue", NA, "brown", "brown", "haz...
## $ hair   <chr> NA, "light brown", "blond", "light brown", NA, "brown",...
## $ race   <chr> "NBNH", "NBNH", "NBNH", "NBNH", "NBNH", "NBNH", "NBNH",...
## $ sex    <chr> "female", "female", "female", "female", "male", "male",...
## $ height <int> 65, 62, NA, 67, 63, 64, 65, 65, 66, 66, 71, 66, 71, 67,...
## $ BMI    <dbl> NA, 21.94843, NA, 17.22855, 23.02862, 34.33015, 21.7996...
```

Next, we set up the conversion table to standardize the income values. This will be used later in the analysis.

```
# Load proportional dollar values.  All multipliers convert to 2014 dollars.  Source:  https://westegg.

income_2014_converter <- data_frame(year = sort(unique(income_data_nlsy79$year))) %>%
  mutate(multiplier = c(2.42, 2.34, 2.24, 2.17, 2.13, 2.05, 1.97, 1.88, 1.78, 1.71, 1.66, 1.61, 1.57, 1
```

Analysis of Race vs. Income

Next, we turn our attention to the physical characteristic of race, investigating what effect, if any, it may have on an individual’s income.

Preparation of the data

The income multipliers loaded earlier are multiplied by the raw income data to produce a “true_income” value that adjusts for inflation and other economic conditions, allowing that the data may be compared from year to year.

```
income_phys_all <- income_data_nlsy79 %>%
  inner_join(physical_data_nlsy79) # %>%
```

```
## Joining, by = c("CASEID", "year")
```

```
  # filter(year == 2014)

income_race_all <- select(income_phys_all, CASEID, income, race, year)
income_race_all <- filter(income_race_all, !is.na(income), !is.na(race), !is.na(year))
income_race_all
```

```
## # A tibble: 163,071 x 4
##    CASEID income race   year
##     <int>  <int> <chr> <int>
##  1      2  10000 NBNH   1982
##  2      3   7000 NBNH   1982
##  3      4   1086 NBNH   1982
##  4      5   2300 NBNH   1982
##  5      6   3250 NBNH   1982
##  6      7   4975 NBNH   1982
##  7      8   7500 NBNH   1982
##  8      9   5000 NBNH   1982
##  9     10   9000 NBNH   1982
## 10     11   4002 NBNH   1982
## # ... with 163,061 more rows
```

```
# Convert income to a standard amount for comparison.  Convert all values to 2014 dollars.
income_race_all <- inner_join(income_race_all, income_2014_converter) %>%
  mutate(true_income = income * multiplier)
```

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

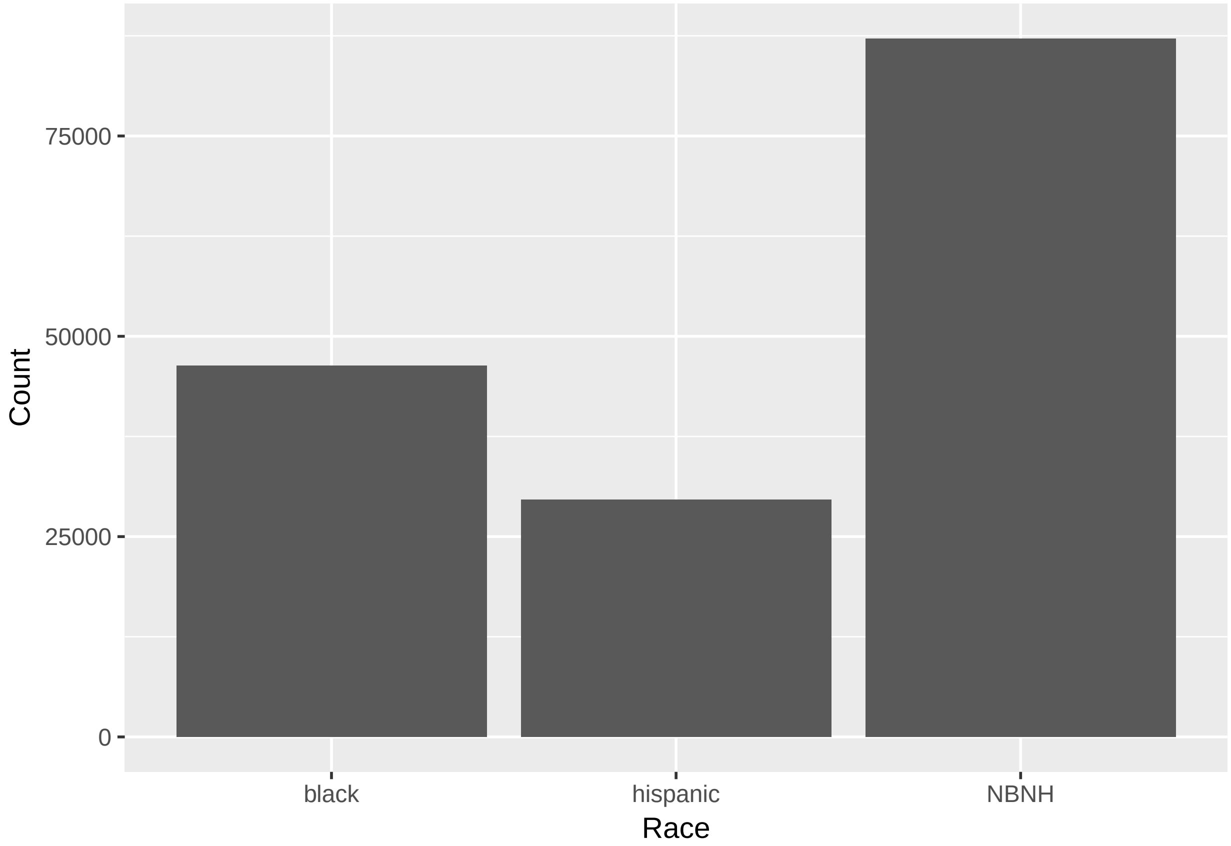
Examination of the race variable

```
unique(income_race_all$race)
```

```
## [1] "NBNH"     "hispanic" "black"
```

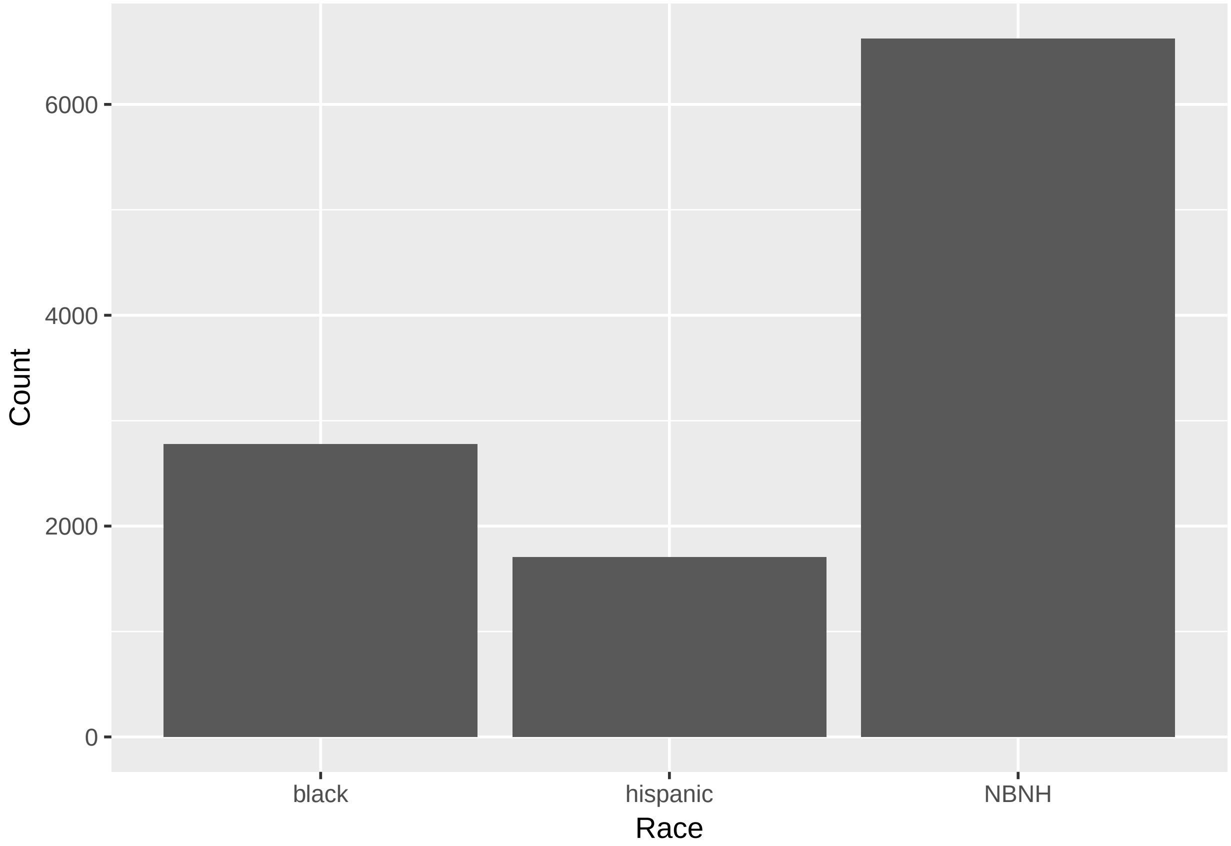
Three values are listed for race, “black”, “hispanic”, and “NBNH” (Neither Black Nor Hispanic). As we will see, the third category contains more records than both of the other categories combined.

```
# Check sample size

ggplot(
  data = income_race_all,
  aes(x = race)
) + geom_bar() +
  scale_y_continuous(name = "Count") +
  scale_x_discrete(name = "Race")
```

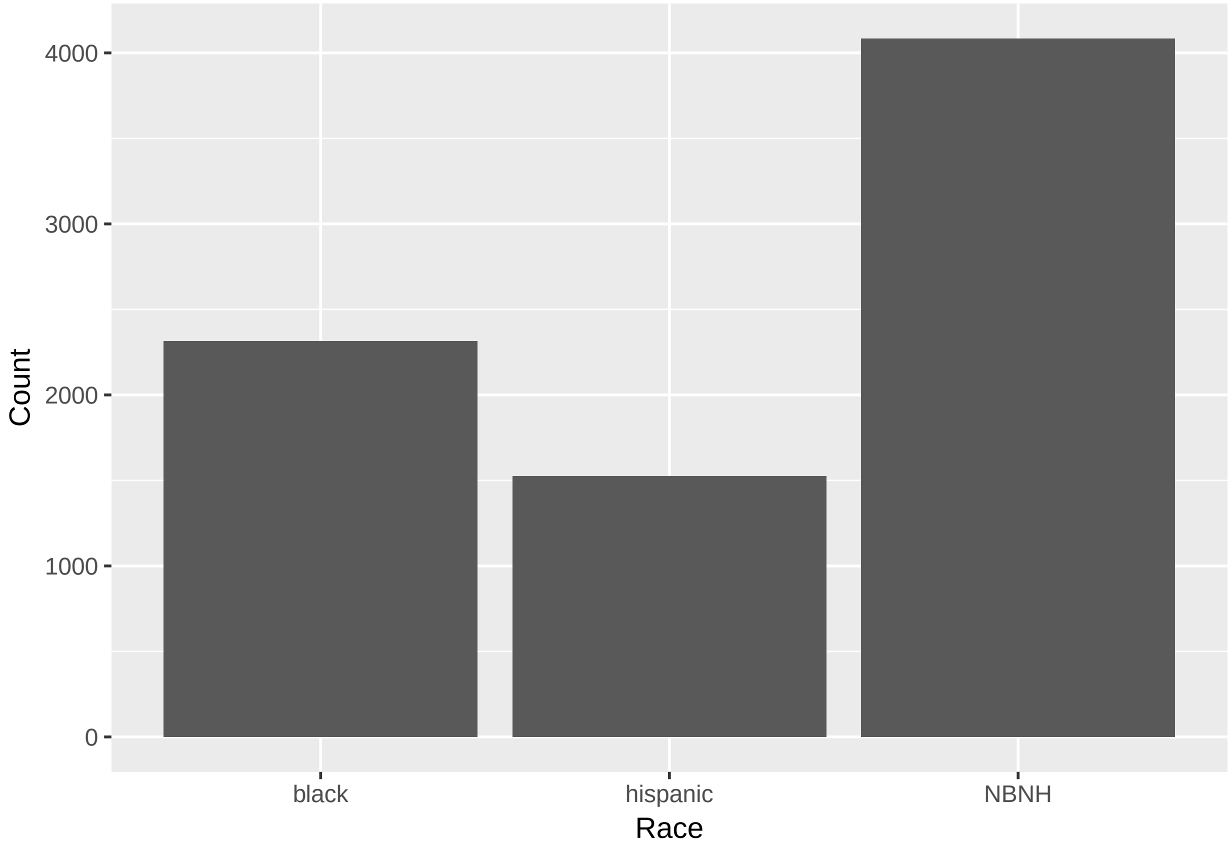


This trend holds true for individual years as well, as seen in 1982, 1998, and 2014 below.

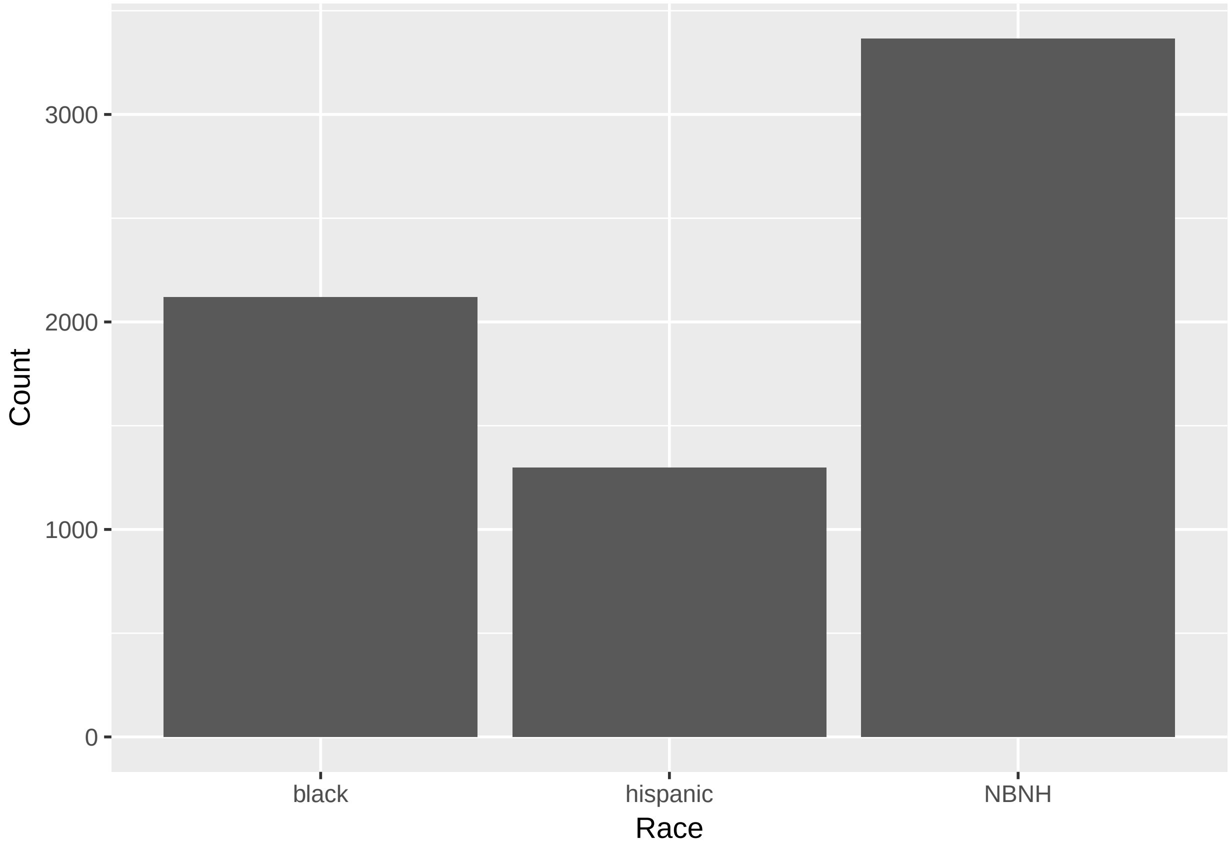
```
# 1982

ggplot(
  data = filter(income_race_all, year == 1982),
  aes(x = race)
) + geom_bar() +
  scale_y_continuous(name = "Count") +
  scale_x_discrete(name = "Race")
```



```
# 1998

ggplot(
  data = filter(income_race_all, year == 1998),
  aes(x = race)
) + geom_bar() +
  scale_y_continuous(name = "Count") +
  scale_x_discrete(name = "Race")
```



```
# 2014

ggplot(
  data = filter(income_race_all, year == 2014),
  aes(x = race)
) + geom_bar() +
  scale_y_continuous(name = "Count") +
  scale_x_discrete(name = "Race")
```



Fortunately, the overall sample size of each individual year and the large sample across all years still allows for meaningful analysis of the black and hispanic race variables.

Examination of income variable

```
head(sort(unique(income_race_all$true_income)))
```

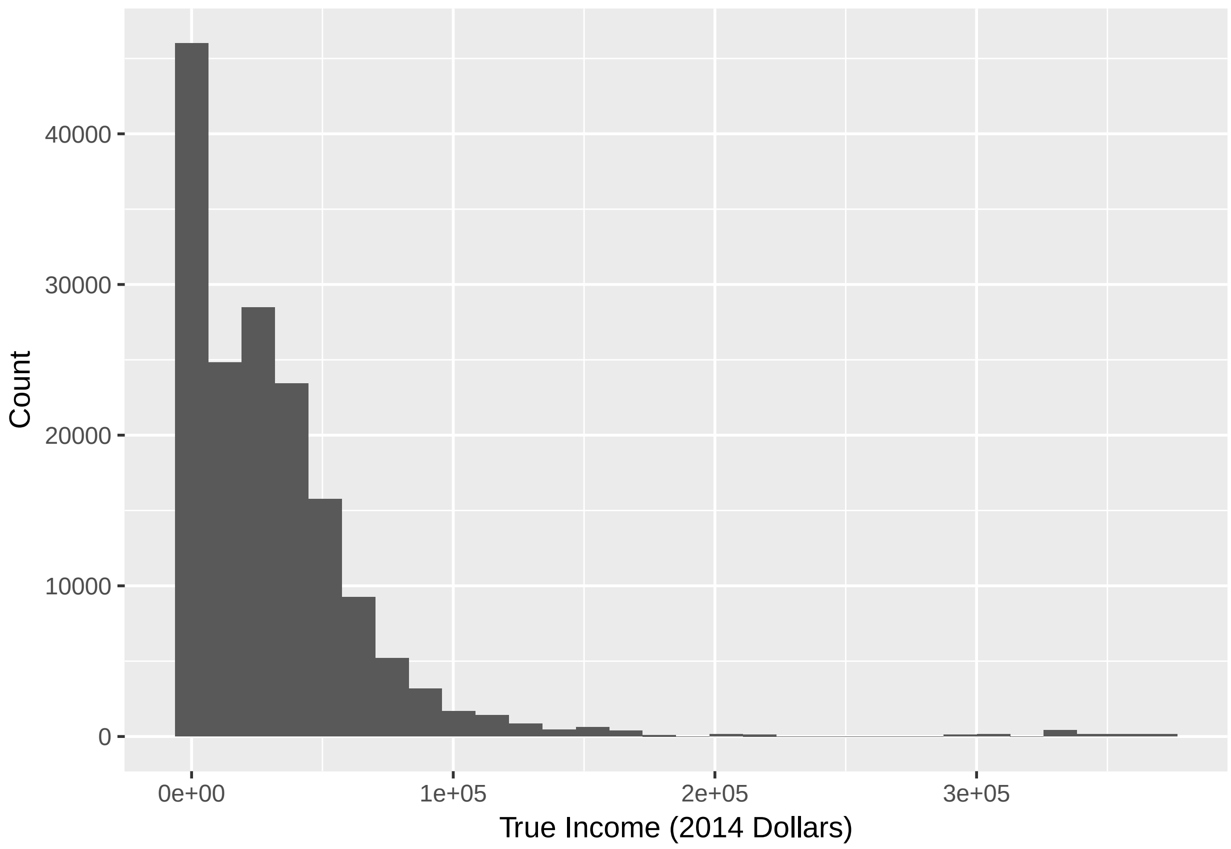
```
## [1] 0.00 1.00 1.10 1.31 1.36 1.61
```

```
head(sort(unique(income_race_all$true_income), decreasing = TRUE))
```

```
## [1] 370314.0 354144.9 338605.3 337309.9 329756.9 327384.7
```

```
ggplot(data = income_race_all, aes(x = income_race_all$true_income)) +
  geom_histogram()+
  scale_y_continuous(name = "Count") +
  scale_x_continuous(name = "True Income (2014 Dollars)")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



High income values

A quick glance at the income data reveals several peak outliers. These are likely the truncation values discussed in the introduction, but we will examine further to determine whether or not they are valid for this analysis.

We start by looking at all values above $200,000 and getting a count of unique values to see if we can isolate the truncation numbers.

```
high_income <- filter(income_race_all, income > 200000)
unique(high_income$income)
```

```
## [1] 216200 236000 265933 279816 307823 312324 343830 370314
```

```
nrow(filter(income_race_all, income == 265933))
```

```
## [1] 133
```

```
nrow(filter(income_race_all, income == 279816))
```

```
## [1] 144
```

```
nrow(filter(income_race_all, income == 307823))
```

```
## [1] 146
```

```
nrow(filter(income_race_all, income == 312324))
```

```
## [1] 140
```

```r
nrow(filter(income_race_all, income == 343830))
```

```
## [1] 143
```

```r
nrow(filter(income_race_all, income == 370314))
```

```
## [1] 144
```

```r
low_income <- filter(income_race_all, income <= 200000)

# sort(unique(income_race_all$true_income))
```

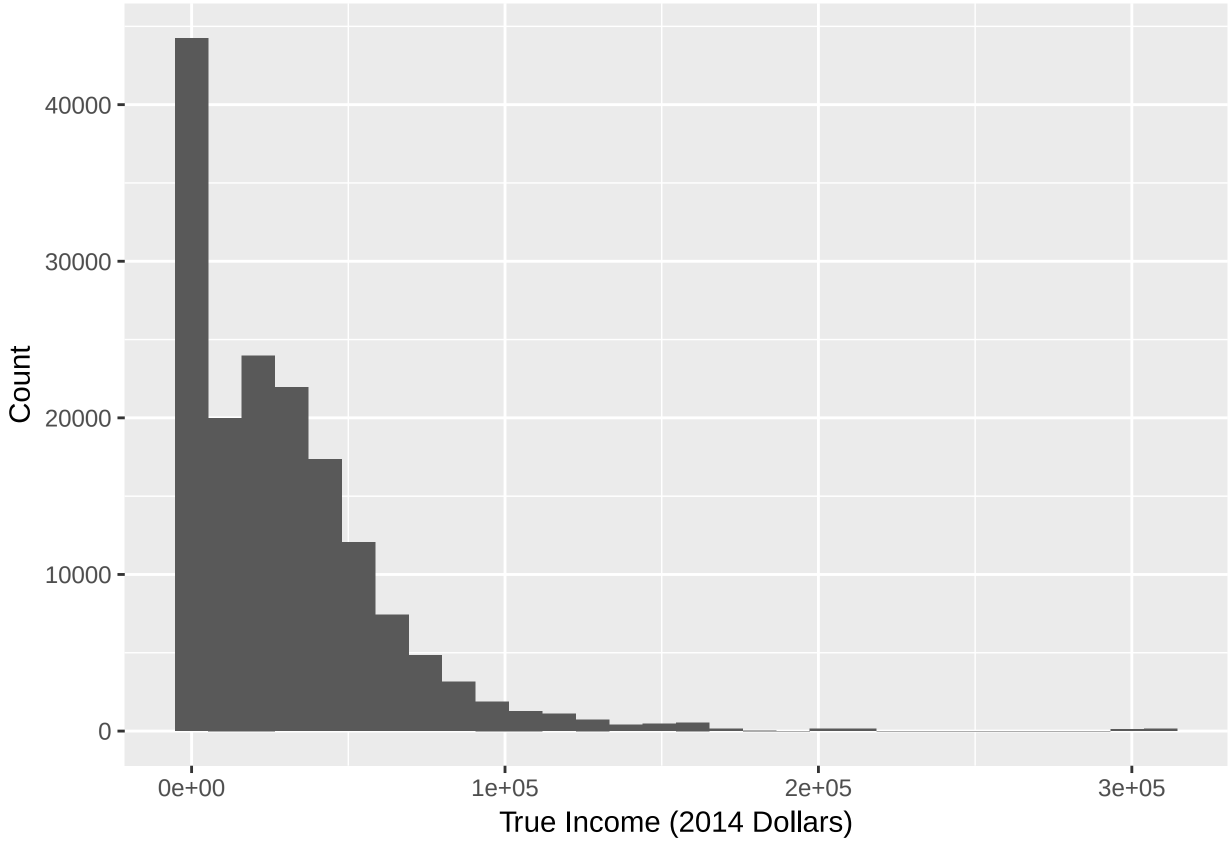
265933 279816 307823 312324 343830 370314 appear to be truncated income values. They are at the top of the scale and between 133 and 146 individuals have the exact same unique income. That's an awfully specific number for so many people to have it.

Looking at the people with incomes below $200,000, the numbers appear much tidier, mostly ending in round hundreds or thousands, but nearly all ending in zeros.

It seems reasonable to exclude these values at the upper end for analysis, given their innacurate nature due to truncation.

```r
income_race_all <- filter(income_race_all, income < 200000 | income %in% c(216200,236000))

ggplot(data = income_race_all, aes(x = income_race_all$true_income)) +
  geom_histogram()+
  scale_y_continuous(name = "Count") +
  scale_x_continuous(name = "True Income (2014 Dollars)")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Zero income values

Next, we look at the peak at $0 income.

```
filter(income_race_all, income == 0)
```

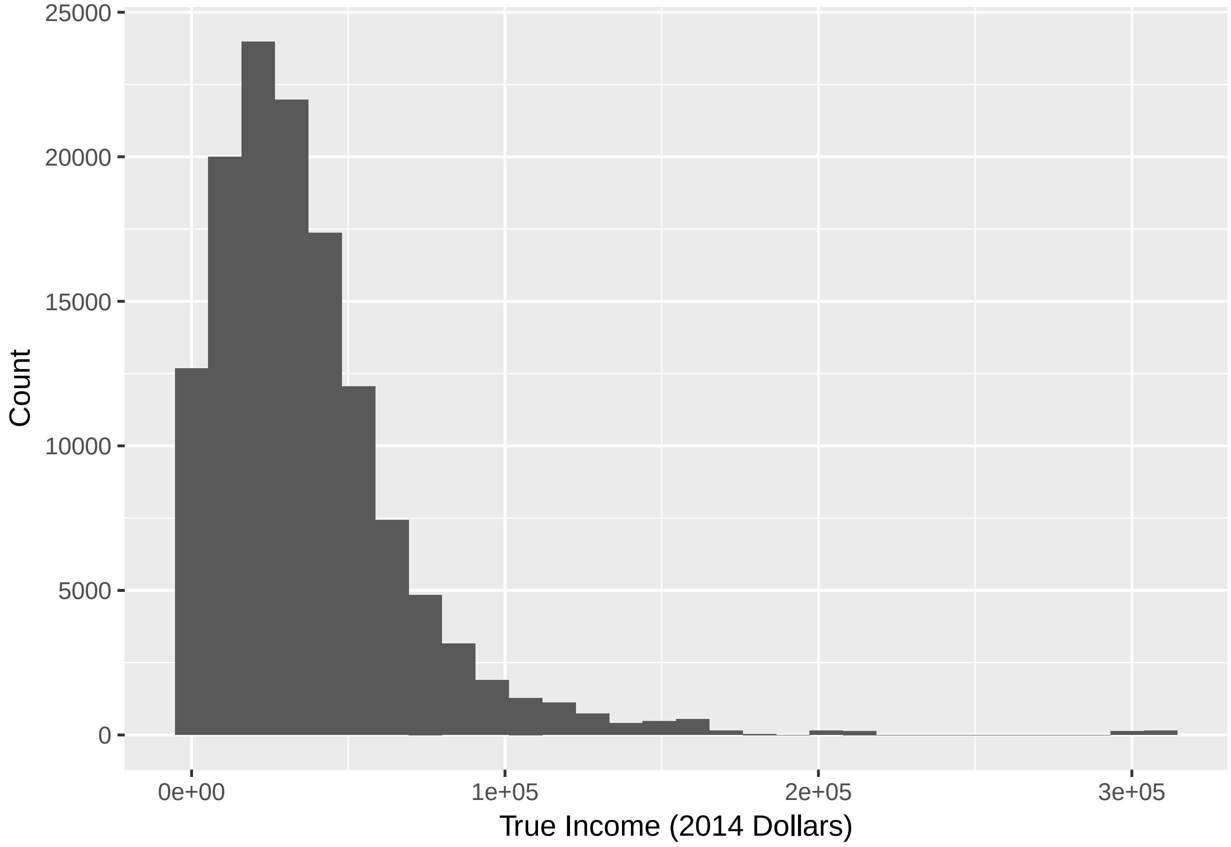
```
## # A tibble: 31,558 x 6
##    CASEID income race      year multiplier true_income
##     <int>  <int> <chr>    <int>      <dbl>       <dbl>
##  1     15      0 NBNH      1982       2.42           0
##  2     31      0 hispanic  1982       2.42           0
##  3     35      0 NBNH      1982       2.42           0
##  4     40      0 NBNH      1982       2.42           0
##  5     86      0 black     1982       2.42           0
##  6     87      0 black     1982       2.42           0
##  7     91      0 NBNH      1982       2.42           0
##  8     98      0 NBNH      1982       2.42           0
##  9    113      0 NBNH      1982       2.42           0
## 10    127      0 NBNH      1982       2.42           0
## # ... with 31,548 more rows
```

31,558 records with 0 for income. These could be due to stay-at-home mothers, individuals in school, unemployment factors unrelated to race, or any other number of possibilities. For the purposes of this analysis, we will exclude the $0 income records.

```
# For now, considering only those records with an income > 0
```

```
income_race_all <- filter(income_race_all, income > 0)

ggplot(data = income_race_all, aes(x = income_race_all$true_income)) +
  geom_histogram() +
  scale_y_continuous(name = "Count") +
  scale_x_continuous(name = "True Income (2014 Dollars)")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We see a clearly left-skewed distribution forming as the data is transformed.

Final validity check

```
# No NA race values
sum(is.na(income_race_all$race))
```

```
## [1] 0
```

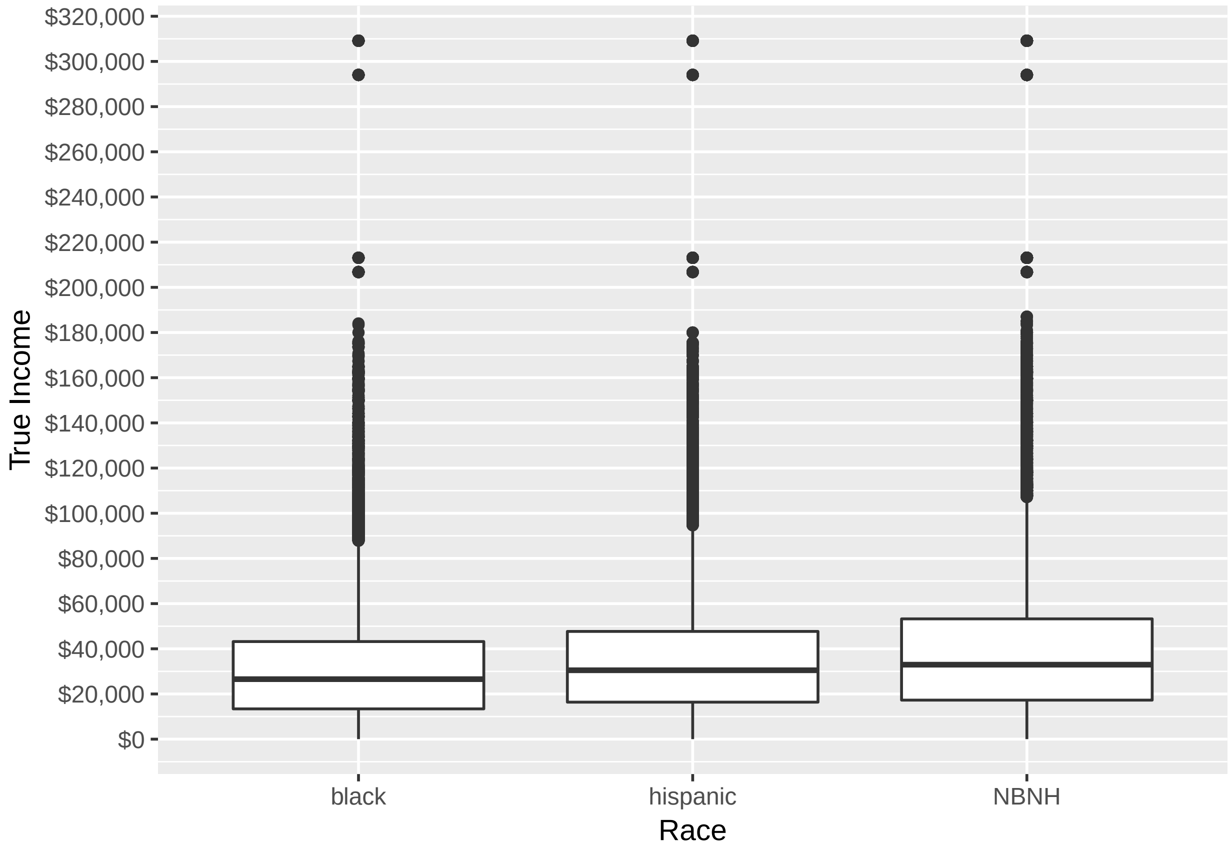
```
sum(is.na(income_race_all$income))
```

```
## [1] 0
```

We see no NA values for race or income, so we are prepared to move ahead with the analysis.

Analysis of Mean and Median incomes by race

```
ggplot(
  data = income_race_all,
  aes(x = race, y = true_income)
) + geom_boxplot() +
    scale_y_continuous(breaks = c(0, 20000, 40000, 60000, 80000,100000, 120000, 140000, 160000, 180000,
    scale_x_discrete(name = "Race")
```



The boxplot above clearly illustrates that the median incomes for black and hispanic individuals are lower than those for other races, and that the spread of incomes generally trends higher for hispanics than blacks, and for other groups more than the two minorities broken out in this data.

Here are the overall mean and median values:

```
print(c("Mean income black: ", mean(income_race_all$true_income[income_race_all$race == "black"])))
```

```
## [1] "Mean income black: " "31496.1686209569"
```

```
print(c("Mean income hispanic: ", mean(income_race_all$true_income[income_race_all$race == "hispanic"])
```

```
## [1] "Mean income hispanic: " "36070.8110082694"
```

```
print(c("Mean income NBNH: ", mean(income_race_all$true_income[income_race_all$race == "NBNH"])))
```

```
## [1] "Mean income NBNH: " "40295.2234527974"
```

```
print(c("Median income black: ", median(income_race_all$true_income[income_race_all$race == "black"]))
## [1] "Median income black: " "26560"
```

```
print(c("Median income hispanic: ", median(income_race_all$true_income[income_race_all$race == "hispanic
```

```
## [1] "Median income hispanic: " "30498.75"
```

```
print(c("Median income NBNH: ", median(income_race_all$true_income[income_race_all$race == "NBNH"])))
```

```
## [1] "Median income NBNH: " "32970"
```

```
print(c("Mean - Median income black: ", mean(income_race_all$true_income[income_race_all$race == "black
```

```
## [1] "Mean - Median income black: " "4936.16862095686"
```

```
print(c("Mean - Median income hispanic: ", mean(income_race_all$true_income[income_race_all$race == "his
```

```
## [1] "Mean - Median income hispanic: " "5572.06100826945"
```

```
print(c("Mean - Median income NBNH: ", mean(income_race_all$true_income[income_race_all$race == "NBNH"])
```

```
## [1] "Mean - Median income NBNH: " "7325.22345279744"
```

```
print(c("Mean - Median proportion of mean black: ", ((mean(income_race_all$true_income[income_race_all$r
```

```
## [1] "Mean - Median proportion of mean black: "
## [2] "0.156722828111621"
```

```
print(c("Mean - Median proportion of mean hispanic: ", ((mean(income_race_all$true_income[income_race_al
```

```
## [1] "Mean - Median proportion of mean hispanic: "
## [2] "0.154475623156685"
```

```
print(c("Mean - Median proportion of mean NBNH: ", ((mean(income_race_all$true_income[income_race_all$ra
```

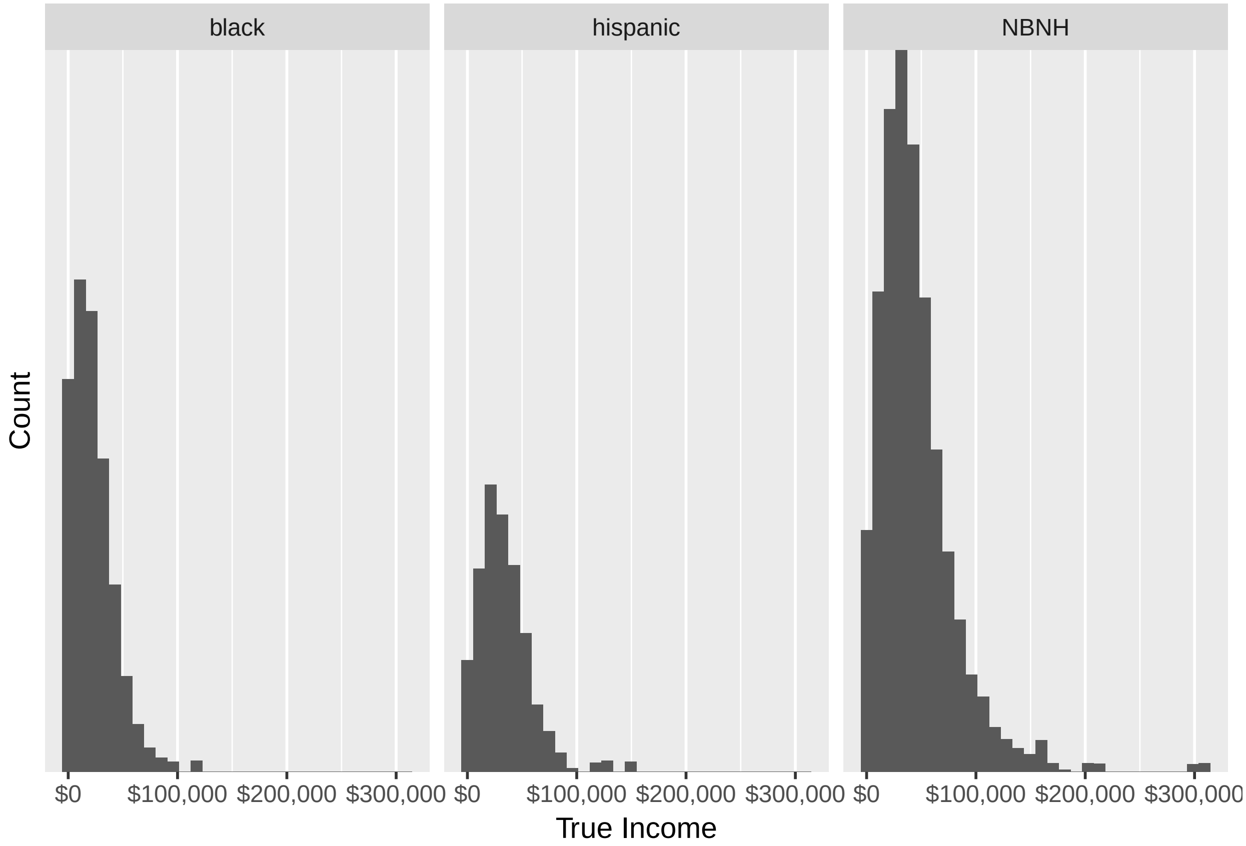
```
## [1] "Mean - Median proportion of mean NBNH: "
## [2] "0.181788877815216"
```

Income spreads by race

Next, let's analyze the income spreads by race.

```
ggplot(data = income_race_all, aes(x = income_race_all$true_income)) +
  geom_histogram() +
  facet_wrap(~race) +
    scale_x_continuous(breaks = c(0, 100000, 200000, 300000,  400000), labels = c("$0", "$100,000", "$20
    scale_y_discrete(name = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



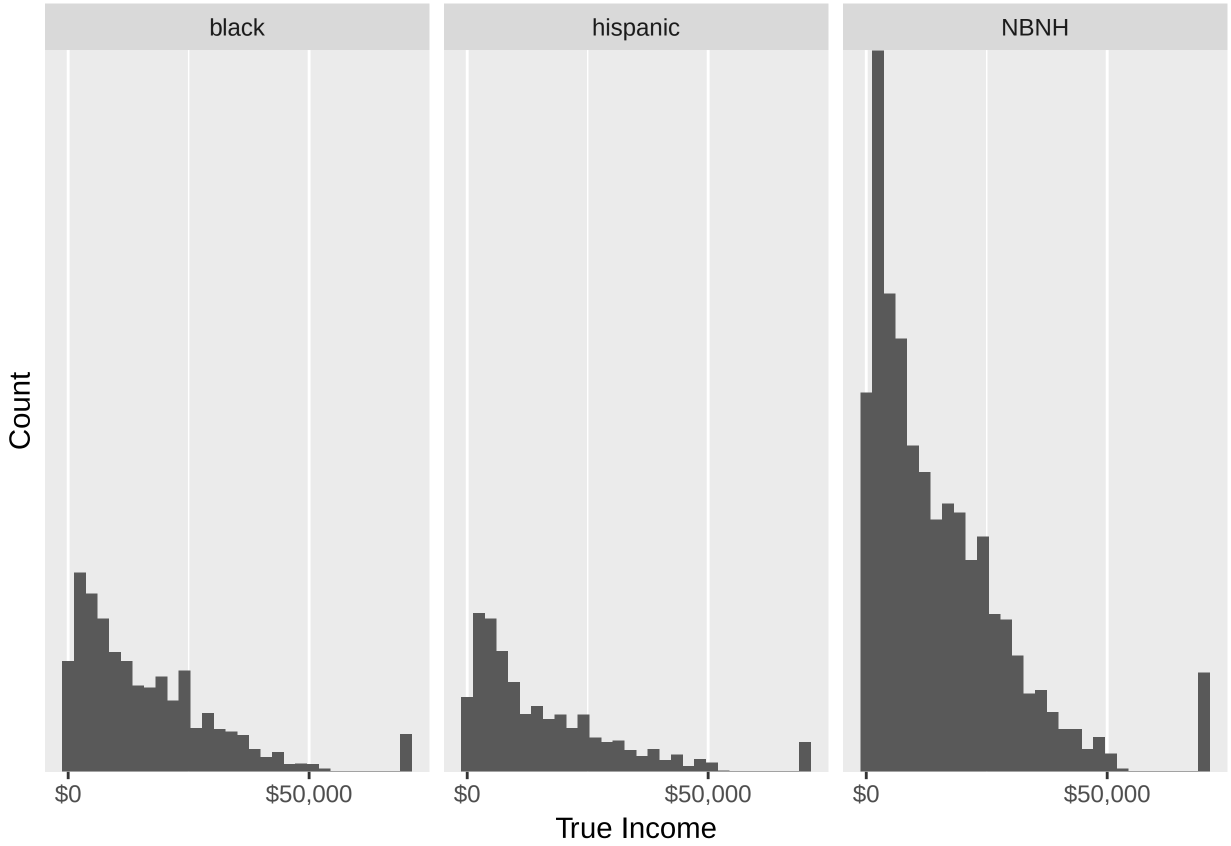
The histogram above shows how the cluster of incomes is skewed further to the right (lower income) for both the black and hispanic subgroups, with black incomes clustering even lower.

Let's break this down into individual years and see if there are any changes.

Income spreads by race, 1982

```
data_1982 <- filter(income_race_all, year == 1982)

ggplot(data = data_1982, aes(x = data_1982$true_income)) +
  geom_histogram() +
  facet_wrap(~race) +
    scale_x_continuous(breaks = c(0, 50000, 100000), labels = c("$0", "$50,000", "$100,000"), name = "T
    scale_y_discrete(name = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

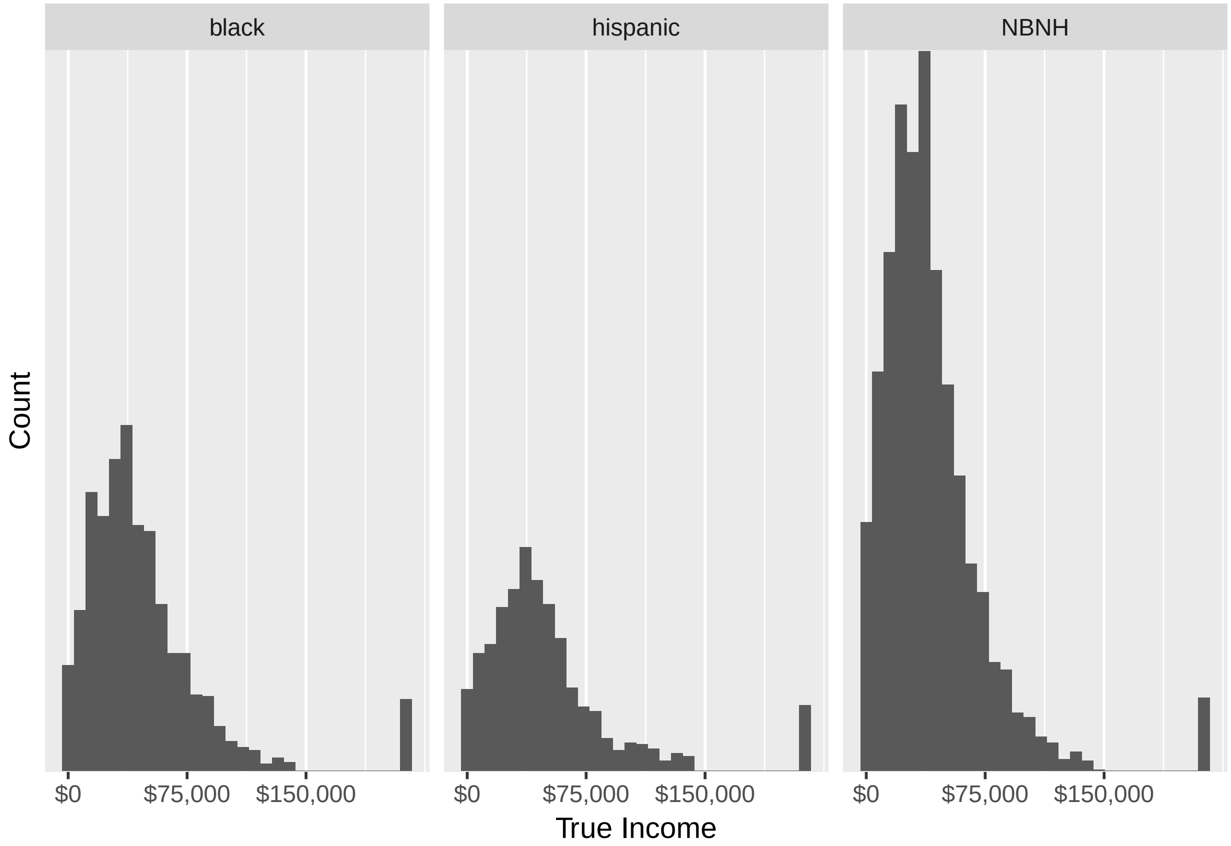


The right-skew of the NBNH group is much clearer in 1982 than for the whole population. Next, we examine 1998, chosen after the enacting of NAFTA to see if there is any different trend.

Income spreads by race, 1998

```
data_1998 <- filter(income_race_all, year == 1998)

ggplot(data = data_1998, aes(x = data_1998$true_income)) +
  geom_histogram() +
  facet_wrap(~race) +
    scale_x_continuous(breaks = c(0, 75000, 150000), labels = c("$0", "$75,000", "$150,000"), name = "T
    scale_y_discrete(name = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



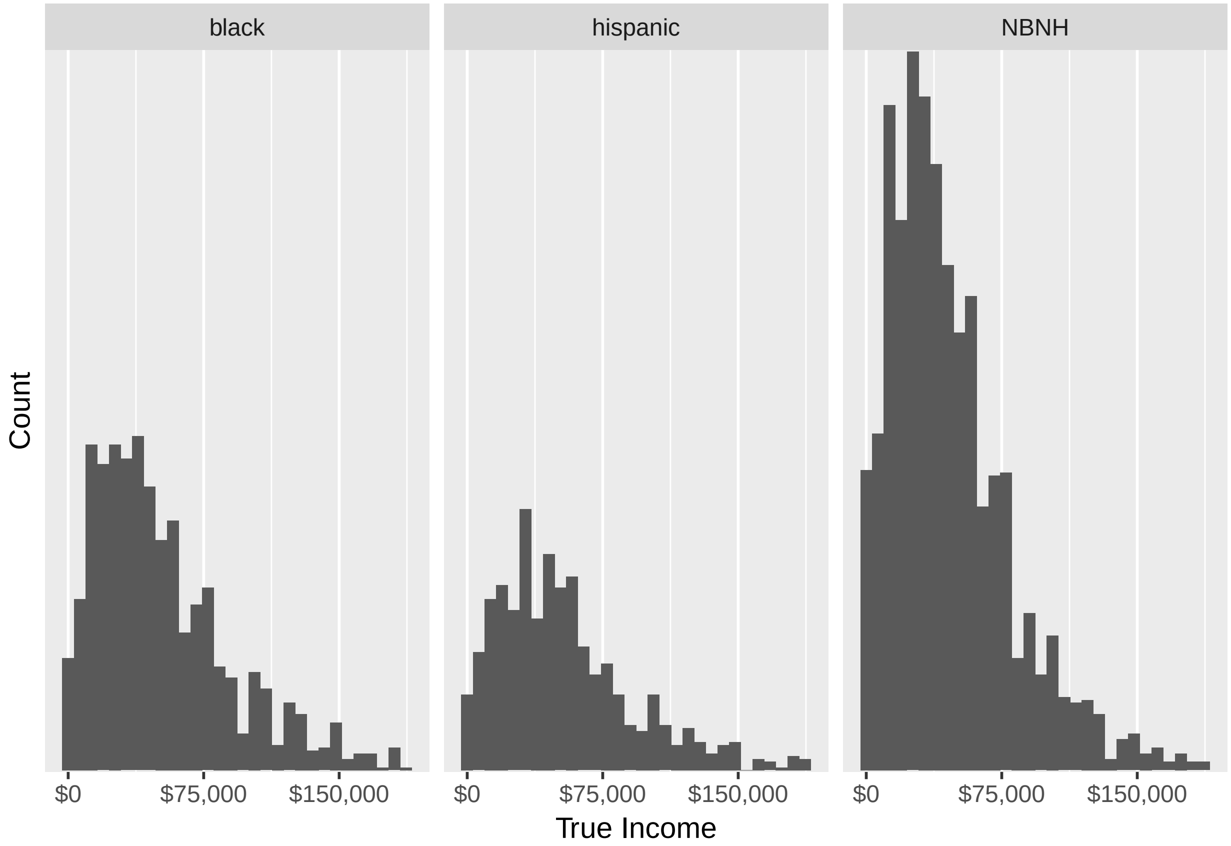
The hispanic spread has clearly shifted to the left. As NAFTA allowed for a more free exchange of goods from Mexico, it is reasonable that the incomes of hispanic people may have seen an increase. This will be a hypothesis for further investigation.

Finally, we examine the last year of the data, 2014.

Income spreads by race, 2014

```
data_2014 <- filter(income_race_all, year == 2014)

ggplot(data = data_2014, aes(x = data_2014$true_income)) +
  geom_histogram() +
  facet_wrap(~race) +
    scale_x_continuous(breaks = c(0, 75000, 150000), labels = c("$0", "$75,000", "$150,000"), name = "T
    scale_y_discrete(name = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



This histogram shows a pronounced leftward shift in the incomes of those individuals listed as black, as compared to the data in 1998 or 1982. However, we also see more outlying peaks in the upper end of the income scale. It could be that a few black high-earners are pulling the mean data up disproportionately compared to the hispanic or NBNH race groups.

Conclusion and Hypotheses

Hypotheses to consider for further analysis (Race vs Income)

1) The proportion of high earners is greater for NBNH than for either group, and for hispanics greater than blacks. Do a few very-high earners skew the numbers disproportionately in both minority groups, but particularly the black group, making the mean values appear more equal than truly represented in the population?

2) Does the hypothesis about NAFTA increasing incomes for hispanic individuals hold? A further analysis of the pre- and post-NAFTA years is required to fully answer this question, but we expect that mean and median incomes increase for the hispanic population as a result of NAFTA, and that this increase comes as the black population declines or grows more slowly.

3) The NBNH group seems to quietly outpace the two minority groups throughout the study. We hypothesize that the gap between NBNH and the black and hispanic groups combined has grown over the course of the study.