Data Cleaning

Sara Elfring

2024-12-03

First Look

```
# Load in NHANES data
data("NHANES")
```

```
# Select Covariates and Summarize
NHANES |>
  select(BPSysAve, Age, BMI, Gender, Alcohol12PlusYr,
         AlcoholDay, AlcoholYear) |>
  summary()
```

```
##     BPSysAve          Age              BMI           Gender     Alcohol12PlusYr
##  Min.   : 76.0   Min.   : 0.00   Min.   :12.88   female:5020   No  :1368
##  1st Qu.:106.0   1st Qu.:17.00   1st Qu.:21.58   male  :4980   Yes :5212
##  Median :116.0   Median :36.00   Median :25.98                 NA's:3420
##  Mean   :118.2   Mean   :36.74   Mean   :26.66
##  3rd Qu.:127.0   3rd Qu.:54.00   3rd Qu.:30.89
##  Max.   :226.0   Max.   :80.00   Max.   :81.25
##  NA's   :1449                    NA's   :366
##    AlcoholDay      AlcoholYear
##  Min.   : 1.000   Min.   :  0.0
##  1st Qu.: 1.000   1st Qu.:  3.0
##  Median : 2.000   Median : 24.0
##  Mean   : 2.914   Mean   : 75.1
##  3rd Qu.: 3.000   3rd Qu.:104.0
##  Max.   :82.000   Max.   :364.0
##  NA's   :5086     NA's   :4078
```

We start with 10000 samples.

Remove samples missing BPSysAve

```
sample = NHANES |>
  select(BPSysAve, Age, BMI, Gender, Alcohol12PlusYr,
         AlcoholDay, AlcoholYear) |>
  filter(!is.na(BPSysAve))
summary(sample)
```

```
##     BPSysAve          Age          BMI          Gender     Alcohol12PlusYr
##  Min.   : 76.0   Min.   : 8   Min.   :12.89   female:4299   No  :1336
##  1st Qu.:106.0   1st Qu.:24   1st Qu.:22.76   male  :4252   Yes :5122
##  Median :116.0   Median :40   Median :26.65                 NA's:2093
##  Mean   :118.2   Mean   :41   Mean   :27.62
##  3rd Qu.:127.0   3rd Qu.:56   3rd Qu.:31.40
##  Max.   :226.0   Max.   :80   Max.   :81.25
##                               NA's   :64
##    AlcoholDay      AlcoholYear
##  Min.   : 1.000   Min.   :  0.00
##  1st Qu.: 1.000   1st Qu.:  3.00
##  Median : 2.000   Median : 24.00
##  Mean   : 2.908   Mean   : 74.68
##  3rd Qu.: 3.000   3rd Qu.:104.00
##  Max.   :82.000   Max.   :364.00
##  NA's   :3724     NA's   :2731
```

There are now 8551 samples.

Remove Samples Missing Alcohol12PlusYr

```
# Removing samples missing Alcohol12PlusYr
sample = sample |>
  filter(!is.na(Alcohol12PlusYr))
summary(sample)
```

```
##     BPSysAve         Age             BMI          Gender     Alcohol12PlusYr
##  Min.   : 78   Min.   :18.00   Min.   :15.02   female:3189   No :1336
##  1st Qu.:110   1st Qu.:32.00   1st Qu.:24.10   male  :3269   Yes:5122
##  Median :119   Median :46.00   Median :27.80
##  Mean   :121   Mean   :47.01   Mean   :28.80
##  3rd Qu.:129   3rd Qu.:60.00   3rd Qu.:32.22
##  Max.   :226   Max.   :80.00   Max.   :81.25
##                                NA's   :43
##    AlcoholDay      AlcoholYear
##  Min.   : 1.000   Min.   :  0.00
##  1st Qu.: 1.000   1st Qu.:  3.00
##  Median : 2.000   Median : 24.00
##  Mean   : 2.909   Mean   : 74.72
##  3rd Qu.: 3.000   3rd Qu.:104.00
##  Max.   :82.000   Max.   :364.00
##  NA's   :1632     NA's   :641
```

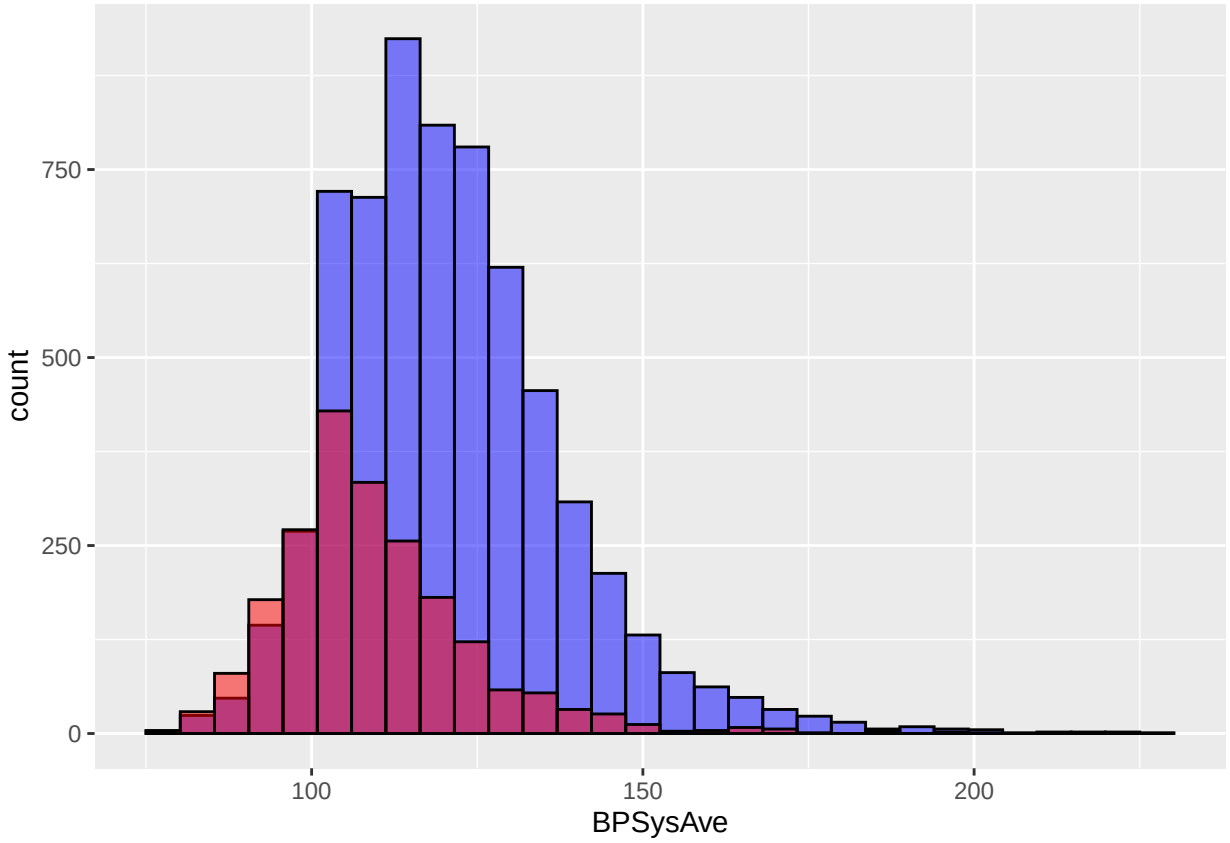
```
# Save the samples missing Alcohol12PlusYr for Comparison
excluded = NHANES |>
  select(BPSysAve, Age, BMI, Gender, Alcohol12PlusYr,
         AlcoholDay, AlcoholYear) |>
  filter(!is.na(BPSysAve), is.na(Alcohol12PlusYr))
```

We have 6458 samples.

Compare the samples missing and not missing Alcohol12PlusYr

```
# Histogram comparison of BPSysAve
ggplot()+
  geom_histogram(data = sample, mapping = aes(x = BPSysAve),
                 fill = "blue", alpha = 0.5, color = "black")+
  geom_histogram(data = excluded, mapping = aes(x = BPSysAve),
                 fill = "red", alpha = 0.5, color = "black")
```

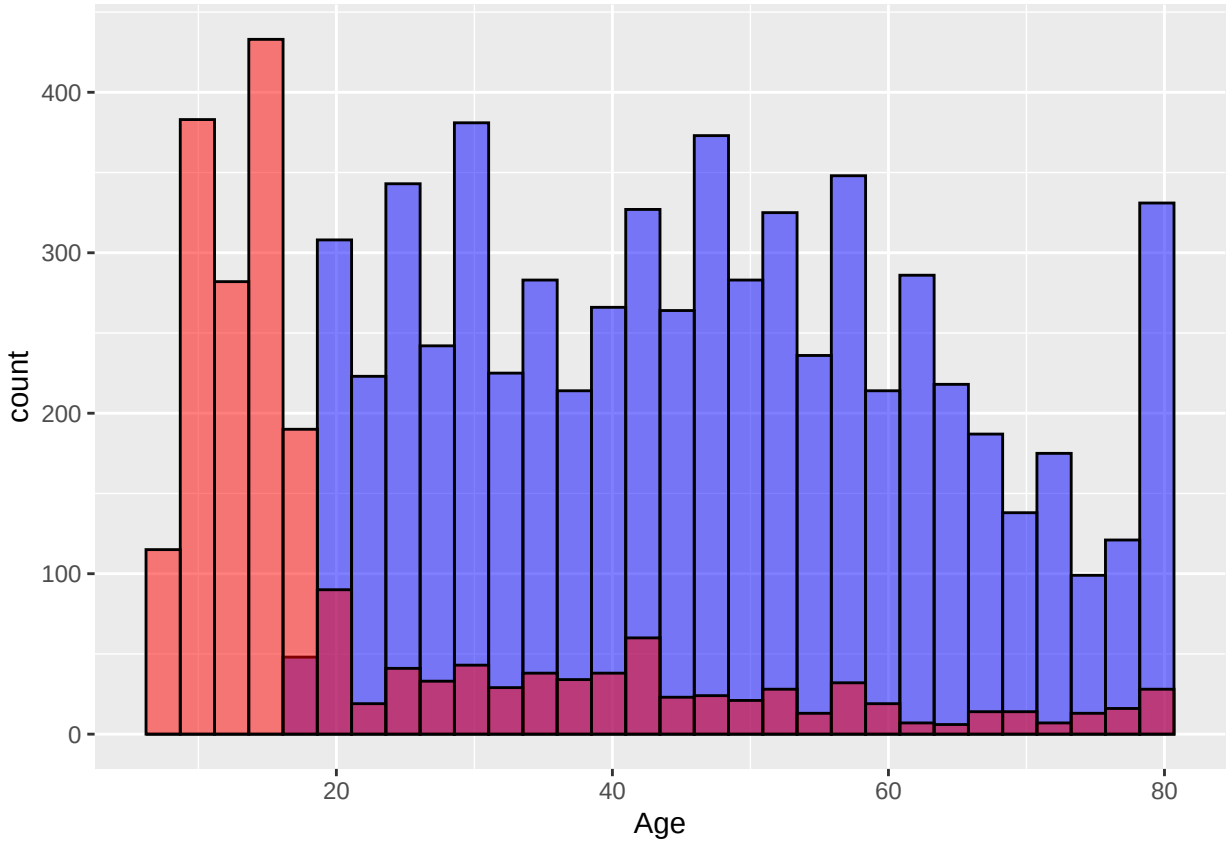
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Samples missing Alcohol12PlusYr have much lower BPSysAve than samples with Alcohol12PlusYr.

```
# Histogram comparison of Age
ggplot()+
  geom_histogram(data = sample, mapping = aes(x = Age),
                 fill = "blue", alpha = 0.5, color = "black")+
  geom_histogram(data = excluded, mapping = aes(x = Age),
                 fill = "red", alpha = 0.5, color = "black")
```

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

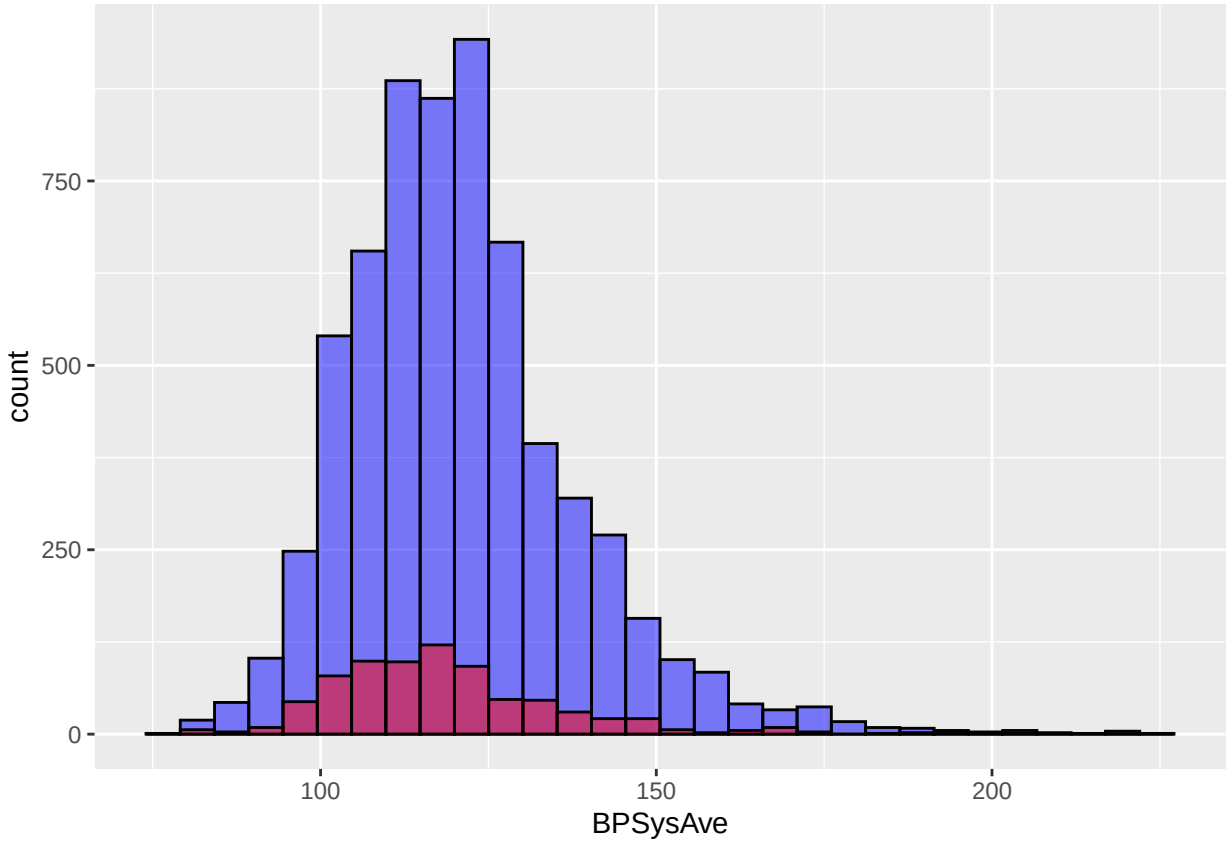


Samples missing Alcohol12PlusYr tend to be younger than samples not missing it. The Alcohol12PlusYr question was only asked of those 18 and older.

```
# Remove samples under 18 years old from the excluded group
excluded = excluded |>
  filter(Age >= 18)
```

```
# Remake histogram comparison of BPSysAve
ggplot()+
  geom_histogram(data = sample, mapping = aes(x = BPSysAve),
                 fill = "blue", alpha = 0.5, color = "black")+
  geom_histogram(data = excluded, mapping = aes(x = BPSysAve),
                 fill = "red", alpha = 0.5, color = "black")
```

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



The distribution of BPSysAve is more similar now between those missing Alcohol12PlusYr and those not missing it.

Remove Samples Missing BMI

```
# Removing samples missing BMI
sample = sample |>
  filter(!is.na(BMI))
summary(sample)
```

```
##     BPSysAve          Age             BMI          Gender     Alcohol12PlusYr
##  Min.   : 78.0   Min.   :18.00   Min.   :15.02   female:3172   No :1325
##  1st Qu.:110.0   1st Qu.:32.00   1st Qu.:24.10   male  :3243   Yes:5090
##  Median :119.0   Median :46.00   Median :27.80
##  Mean   :120.9   Mean   :46.94   Mean   :28.80
##  3rd Qu.:129.0   3rd Qu.:60.00   3rd Qu.:32.22
##  Max.   :226.0   Max.   :80.00   Max.   :81.25
##
##    AlcoholDay      AlcoholYear
##  Min.   : 1.000   Min.   :  0.00
##  1st Qu.: 1.000   1st Qu.:  3.00
##  Median : 2.000   Median : 24.00
##  Mean   : 2.912   Mean   : 74.82
##  3rd Qu.: 3.000   3rd Qu.:104.00
```

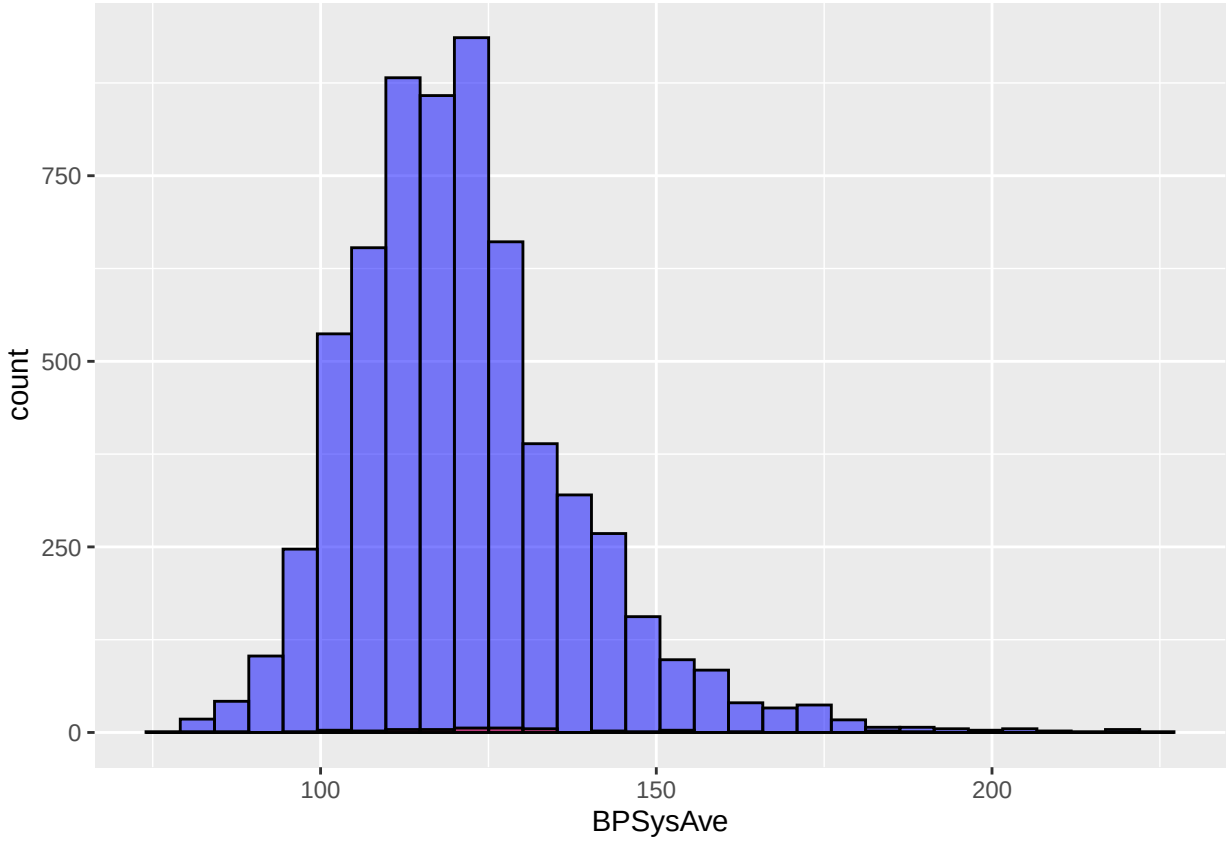
```
##  Max.   :82.000   Max.   :364.00
##  NA's   :1611     NA's   :632
```

```
# Save the samples missing BMI for Comparison
excluded = NHANES |>
  select(BPSysAve, Age, BMI, Gender, Alcohol12PlusYr,
         AlcoholDay, AlcoholYear) |>
  filter(!is.na(BPSysAve), !is.na(Alcohol12PlusYr), is.na(BMI))
```

We have 6415 samples.

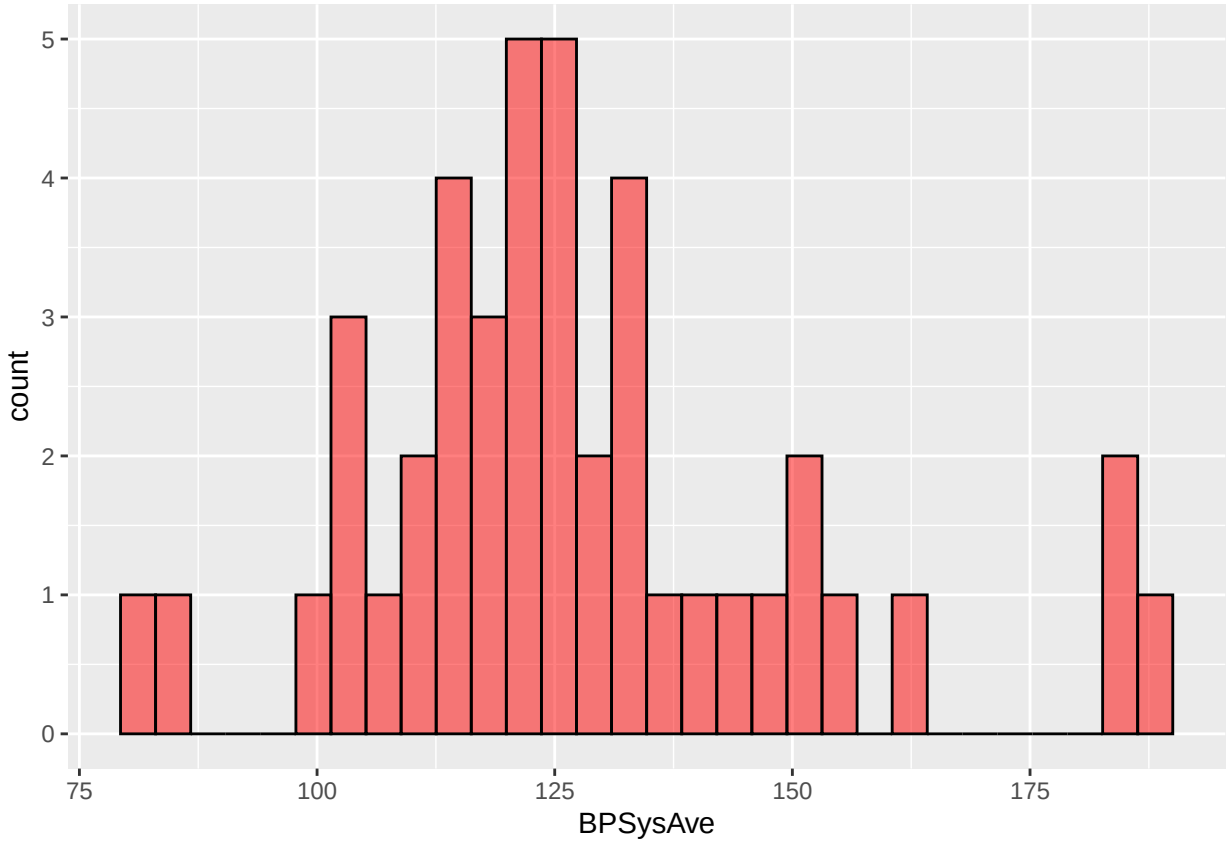
```
# Histogram comparison of BPSysAve
ggplot()+
  geom_histogram(data = sample, mapping = aes(x = BPSysAve),
                 fill = "blue", alpha = 0.5, color = "black")+
  geom_histogram(data = excluded, mapping = aes(x = BPSysAve),
                 fill = "red", alpha = 0.5, color = "black")
```

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




```
ggplot()+
  geom_histogram(data = excluded, mapping = aes(x = BPSysAve),
                 fill = "red", alpha = 0.5, color = "black")
```

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



The BPSysAve values for samples missing BMI fall in the middle of the distribution of BPSysAve for samples not missing BMI.

Remove Samples Missing AlcoholDay

```
# Removing samples missing AlcoholDay
sample = sample |>
  filter(!is.na(AlcoholDay))
summary(sample)
```

```
##     BPSysAve          Age             BMI          Gender     Alcohol12PlusYr
##  Min.   : 80.0   Min.   :18.00   Min.   :15.02   female:2235   No : 372
##  1st Qu.:109.0   1st Qu.:31.00   1st Qu.:23.90   male  :2569   Yes:4432
##  Median :118.0   Median :44.00   Median :27.40
##  Mean   :120.2   Mean   :45.12   Mean   :28.38
##  3rd Qu.:129.0   3rd Qu.:57.00   3rd Qu.:31.68
##  Max.   :221.0   Max.   :80.00   Max.   :69.00
##
##    AlcoholDay      AlcoholYear
##  Min.   : 1.000   Min.   :  1.00
##  1st Qu.: 1.000   1st Qu.: 12.00
##  Median : 2.000   Median : 52.00
```

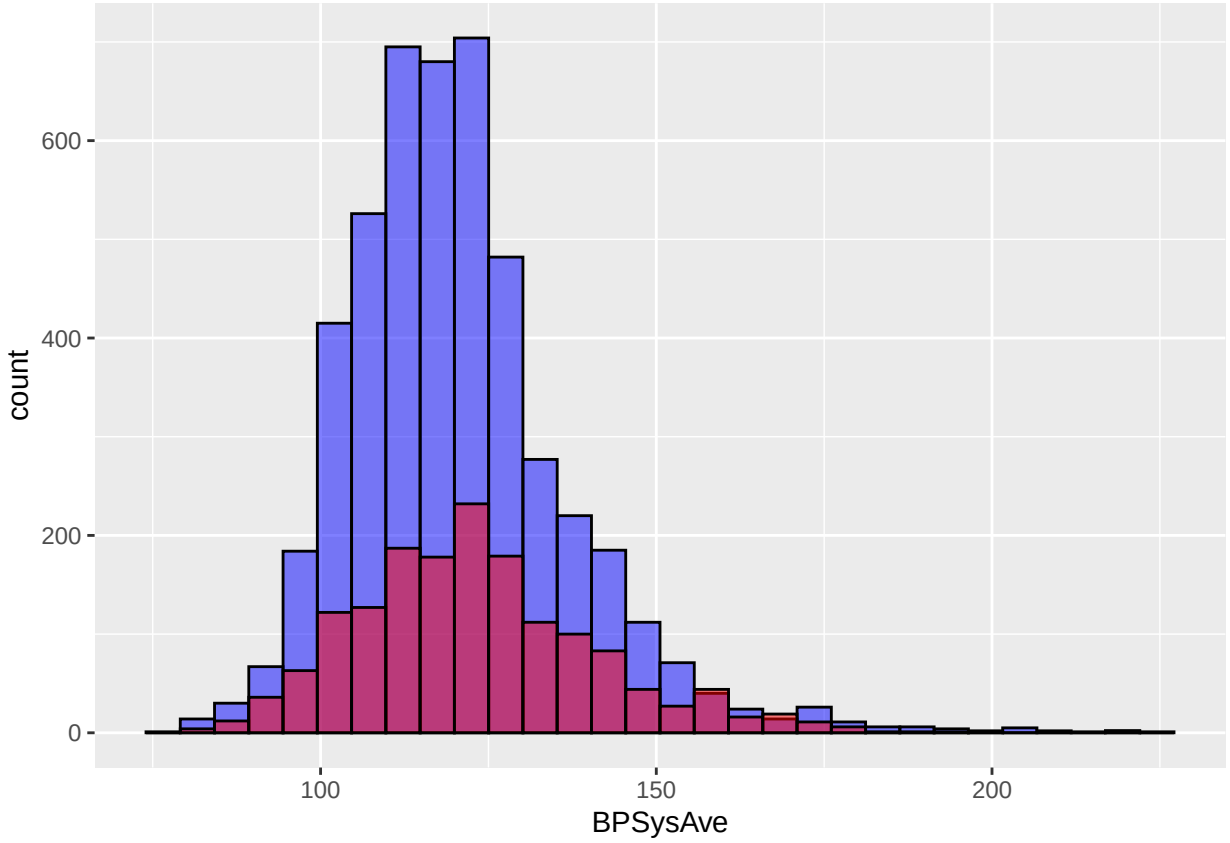
```
##  Mean   : 2.912   Mean   : 90.08
##  3rd Qu.: 3.000   3rd Qu.:156.00
##  Max.   :82.000   Max.   :364.00
##                   NA's   :1
```

```
# Save the samples missing AlcoholDay for Comparison
excluded = NHANES |>
  select(BPSysAve, Age, BMI, Gender, Alcohol12PlusYr,
         AlcoholDay, AlcoholYear) |>
  filter(!is.na(BPSysAve), !is.na(Alcohol12PlusYr), !is.na(BMI),
         is.na(AlcoholDay))
```

We have 4804 samples.

```
# Histogram comparison of BPSysAve
ggplot()+
  geom_histogram(data = sample, mapping = aes(x = BPSysAve),
                 fill = "blue", alpha = 0.5, color = "black")+
  geom_histogram(data = excluded, mapping = aes(x = BPSysAve),
                 fill = "red", alpha = 0.5, color = "black")
```

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



The distribution of BPSysAve is similar now between those missing AlcoholDay and those not missing it.

Remove Samples Missing AlcoholYear

```
# Removing samples missing AlcoholYear
sample = sample |>
  filter(!is.na(AlcoholYear))
summary(sample)
```

```
##     BPSysAve          Age             BMI           Gender     Alcohol12PlusYr
##  Min.   : 80.0   Min.   :18.00   Min.   :15.02   female:2234   No : 372
##  1st Qu.:109.0   1st Qu.:31.00   1st Qu.:23.90   male  :2569   Yes:4431
##  Median :118.0   Median :44.00   Median :27.40
##  Mean   :120.2   Mean   :45.13   Mean   :28.38
##  3rd Qu.:129.0   3rd Qu.:57.00   3rd Qu.:31.68
##  Max.   :221.0   Max.   :80.00   Max.   :69.00
##    AlcoholDay      AlcoholYear
##  Min.   : 1.000   Min.   :  1.00
##  1st Qu.: 1.000   1st Qu.: 12.00
##  Median : 2.000   Median : 52.00
##  Mean   : 2.912   Mean   : 90.08
##  3rd Qu.: 3.000   3rd Qu.:156.00
##  Max.   :82.000   Max.   :364.00
```

There is only one observation missing AlcoholYear removed at this point. We have 4803 samples.

Summary of Our Sample

```
# Number of samples
nrow(sample)
```

```
## [1] 4803
```

```
# Summarize continuous confounders
sample |>
  summarize(mean(BPSysAve), sd(BPSysAve), mean(Age), sd(Age),
            mean(BMI), sd(BMI))
```

```
## # A tibble: 1 x 6
##   `mean(BPSysAve)` `sd(BPSysAve)` `mean(Age)` `sd(Age)` `mean(BMI)` `sd(BMI)`
##              <dbl>          <dbl>       <dbl>     <dbl>       <dbl>     <dbl>
## 1             120.           16.4        45.1      16.5        28.4      6.40
```

```
# Summarize continuous alcohol variables
sample |>
  summarize(mean(AlcoholDay), sd(AlcoholDay),
            mean(AlcoholYear), sd(AlcoholYear))
```

```
## # A tibble: 1 x 4
##   `mean(AlcoholDay)` `sd(AlcoholDay)` `mean(AlcoholYear)` `sd(AlcoholYear)`
##                <dbl>            <dbl>               <dbl>             <dbl>
## 1               2.91             3.19                90.1              106.
```

```r
# Summarize categorical variables
sample |>
  summarize(prop_male = sum(Gender == 'male')/length(Gender),
            prop_drink =
              sum(Alcohol12PlusYr == 'Yes')/length(Alcohol12PlusYr))
```

```
## # A tibble: 1 x 2
##   prop_male prop_drink
##       <dbl>      <dbl>
## 1     0.535      0.923
```