02_epic

Randy

1/4/2022

1. filter the data with height percentage in 0 to 99.99%
2. visit number larger than 10

```
epic_hw0 <- here::here("data", "epic", "Reg_Encounters.csv") %>%
  read_csv(show_col_types = FALSE) %>%
  janitor::clean_names() %>%
  dplyr::select(id = cffidno,
                age = visit_age,
                ht, htpct,
                wt, wtpct) %>%
  na.omit() %>%
  # Tue Jan 04 11:14:02 2022 ------------------------------
  ## remove partial incorrect observations
  ## keep the individuals
  filter(htpct < 99.99 & htpct > 0,
         ## starting at least from 3
         age >= 3)

# names(epic)
## currently has 1710 individuals
## used to be
epic_ind0 <- epic_hw0 %>%
  group_by(id) %>%
  summarize(hmean = mean(ht, na.rm = T),
            hmed = median(ht, na.rm = T),
            wmean = mean(wt, na.rm = T),
            wmed = median(wt, na.rm = T),
            ## this is the time::starting age
            age_min = min(age),
            age_max = max(age),
            age_med = median(age),
            vnum = n()) %>%
  mutate(age_diff = age_max - age_min) %>%
  # Tue Jan 04 09:34:32 2022 ------------------------------
  ## after this step 1761 individuals
  ## with 76497 observations
  ## only select the visit number over 10 times
  filter(vnum >= 10)
  ## after this step is 1664 individuals
  ## 75959 observations

# View(epic_ind)
```

```
# nrow(epic_ind)
# sum(epic_ind$vnum)
# nrow(epic_hw)
```

3. minimal age smaller than 4 3*. age difference larger than 5???

```
id_age4 <- epic_ind0 %>%
  filter(age_min <= 4) %>%
  # filter(age_min <= 5) %>%
  ## filter with age4 1370 individuals
  ## filter with age5 1446 individuals
  dplyr::select(id) %>%
  unlist()

# length(id_age4)

epic_hw1 <- epic_hw0 %>%
  filter(id %in% id_age4)

epic_ind1 <- epic_hw1 %>%
  group_by(id) %>%
  summarize(hmean = mean(ht, na.rm = T),
            hmed = median(ht, na.rm = T),
            wmean = mean(wt, na.rm = T),
            wmed = median(wt, na.rm = T),
            ## this is the time::starting age
            age_min = min(age),
            age_max = max(age),
            age_med = median(age),
            vnum = n()) %>%
  mutate(age_diff = age_max - age_min) %>%
  filter(age_diff >= 5)
  # Tue Jan 04 09:34:32 2022 ------------------------------
  ## only select the visit number over 10 times
  ## 1325 individuals left so far


# Tue Apr 12 11:28:07 2022 ------------------------------
# nrow(epic_ind1)
  # Tue Apr 12 08:23:10 2022 ------------------------------
  ## age difference large than 5
  # age_diff >= 5
  ## 1272 individuals
```

```
epic_ind1 %>% filter(age_diff >= 5) %>% nrow() ## 1272
```

```
## [1] 1272
```

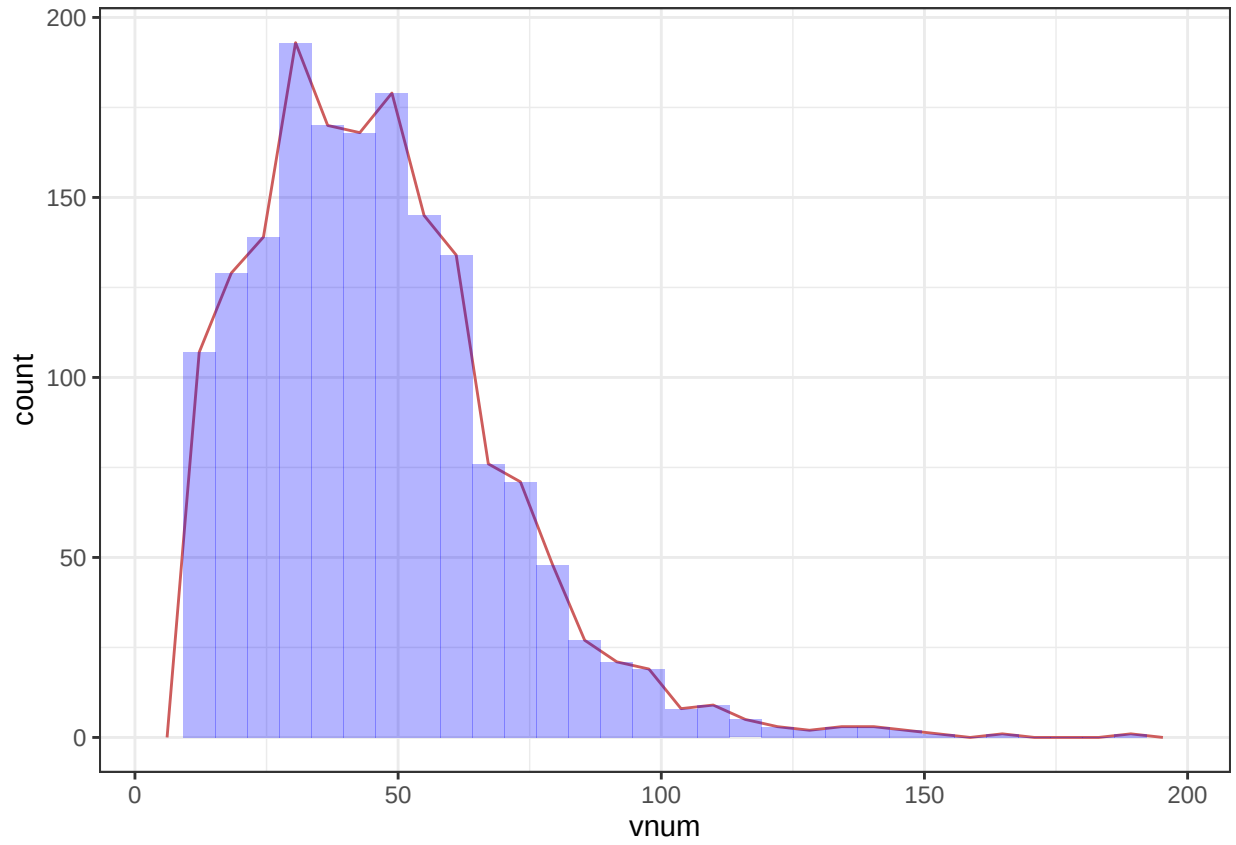
```
epic_ind1 %>% filter(age_diff >= 8) %>% nrow() ## 919
```

```
## [1] 919
```

```
epic10_id <- epic_ind1 %>%
  filter(age_diff >= 10) %>%
  dplyr::select(id, )
  ## 645
```

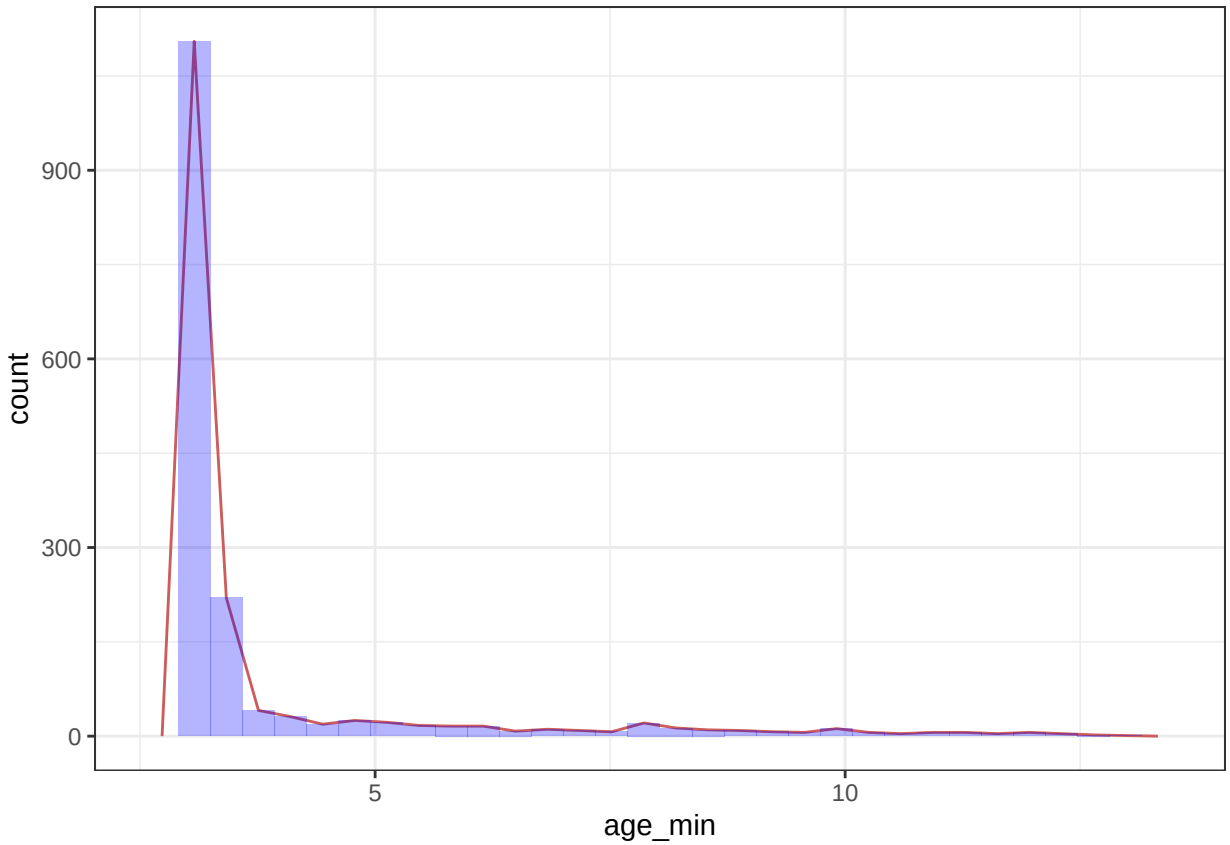
```
# View(epic_10_id)
# summary(epic_ind)

ggplot(data = epic_ind0, aes(vnum)) +
  geom_freqpoly(color = "indianred") +
  geom_histogram(fill = "blue", alpha = 0.3) +
  theme_bw()
```

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



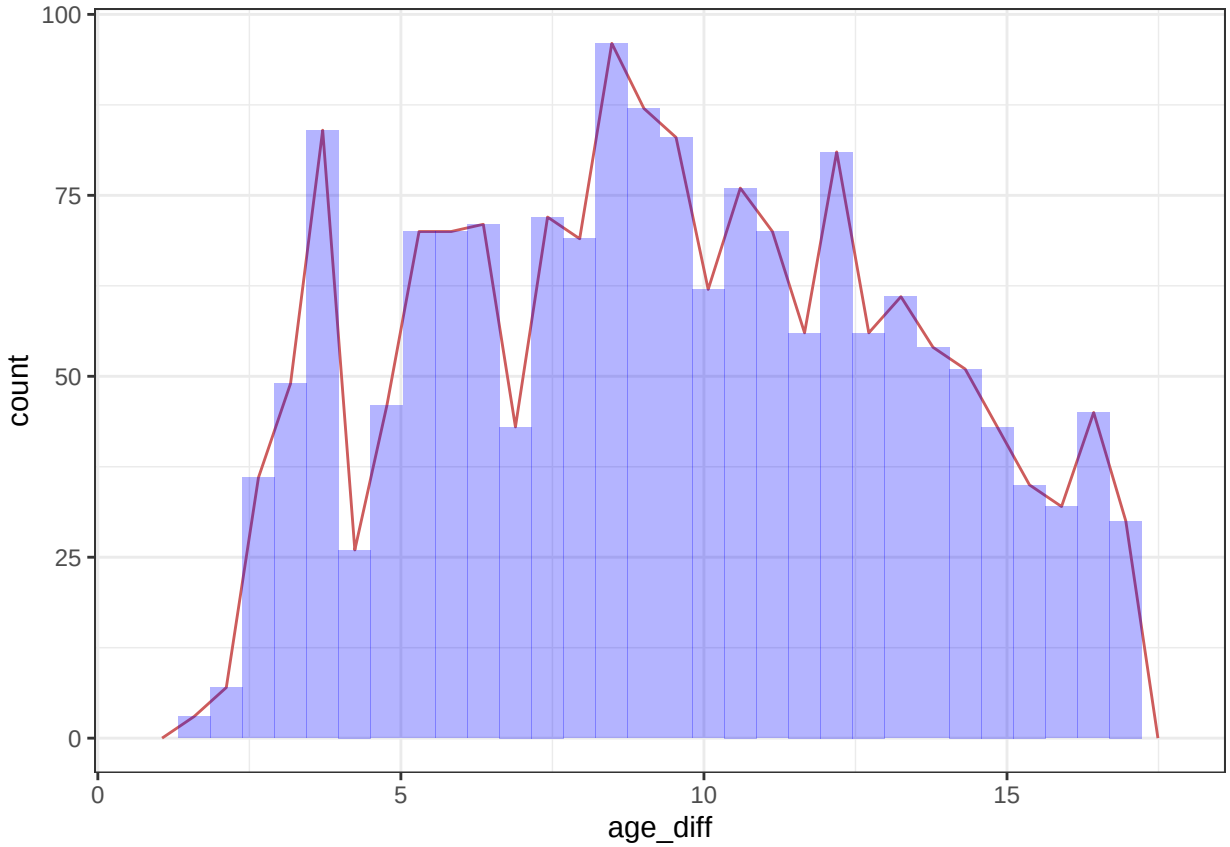

```
ggplot(data = epic_ind0, aes(age_min)) +
  geom_freqpoly(color = "indianred") +
  geom_histogram(fill = "blue", alpha = 0.3) +
  theme_bw()
```

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



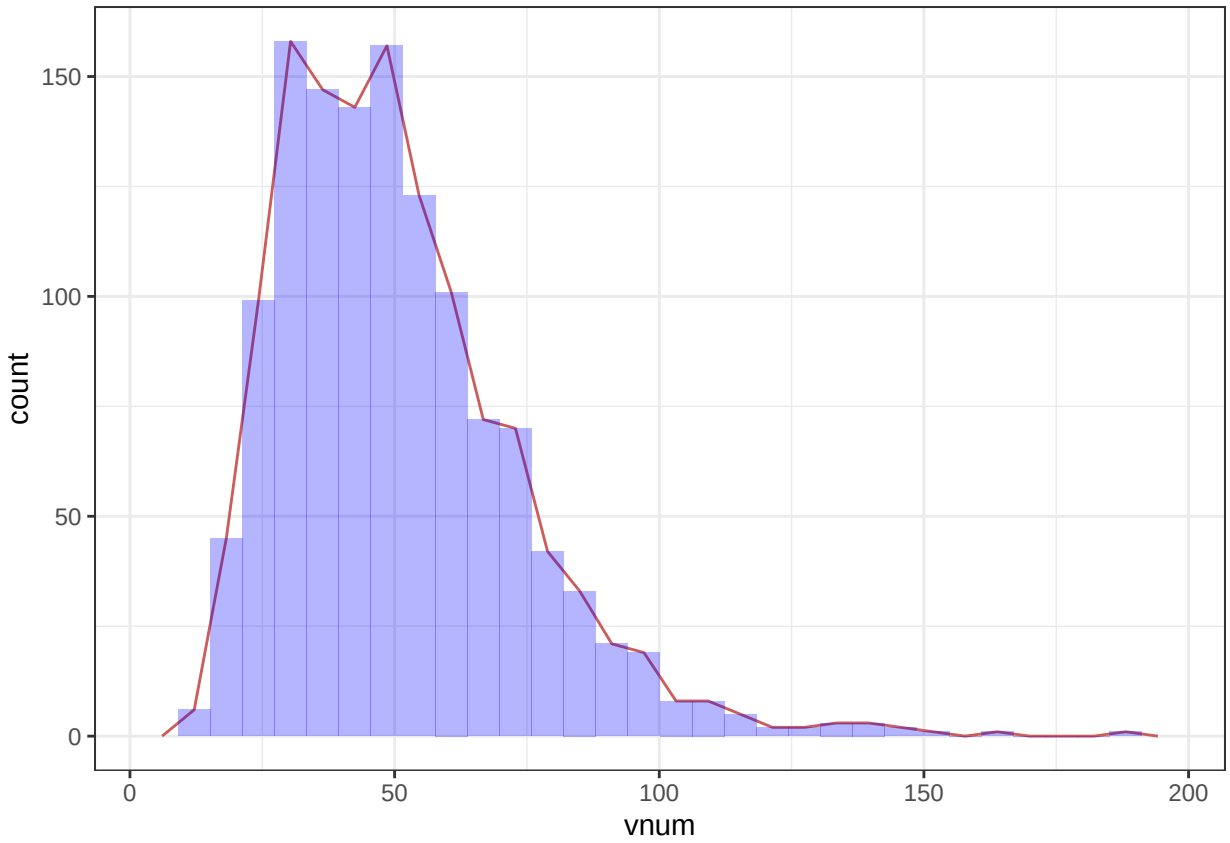

```
ggplot(data = epic_ind0, aes(age_diff)) +
  geom_freqpoly(color = "indianred") +
  geom_histogram(fill = "blue", alpha = 0.3) +
  theme_bw()
```

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



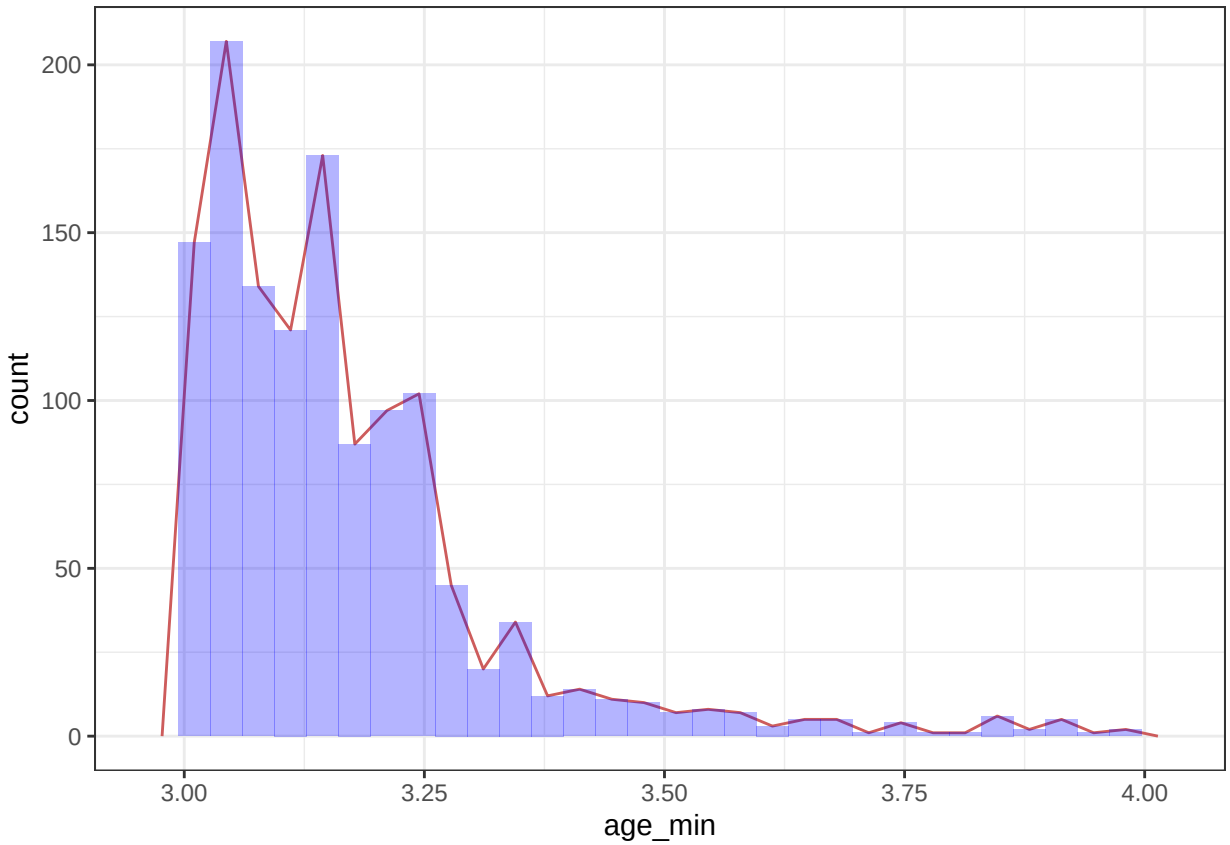

```
ggplot(data = epic_ind1, aes(vnum)) +
  geom_freqpoly(color = "indianred") +
  geom_histogram(fill = "blue", alpha = 0.3) +
  theme_bw()
```

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



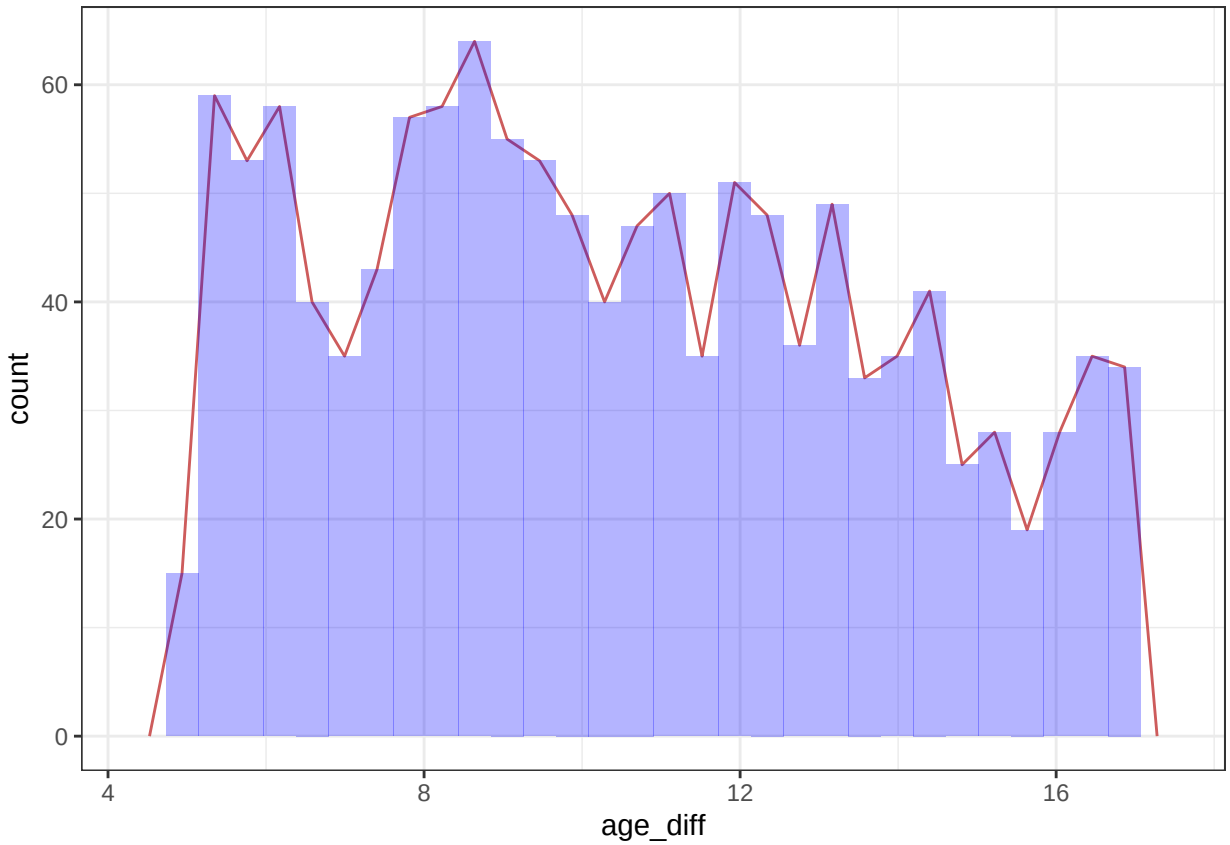

```
ggplot(data = epic_ind1, aes(age_min)) +
  geom_freqpoly(color = "indianred") +
  geom_histogram(fill = "blue", alpha = 0.3) +
  theme_bw()
```

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




```
ggplot(data = epic_ind1, aes(age_diff)) +
  geom_freqpoly(color = "indianred") +
  geom_histogram(fill = "blue", alpha = 0.3) +
  theme_bw()
```

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




```
# hist(epic_ind1$vnum, breaks = 100)
# hist(epic_ind1$age_min, breaks = 40)
# hist(epic_ind1$age_diff, breaks = 40)
# View(epic_hw)
# View(epic_hw)

nrow(epic_hw1) ## 66188
```

```
## [1] 66188
```

```
nrow(epic_ind1) ## 1370
```

```
## [1] 1272
```

```
summary(epic_ind1)
```

```
##        id             hmean           hmed          wmean
##  Min.   :103104   Min.   :102.7   Min.   :101.4   Min.   :16.21
##  1st Qu.:142758   1st Qu.:118.5   1st Qu.:119.1   1st Qu.:23.01
##  Median :152449   Median :126.5   Median :127.5   Median :27.71
##  Mean   :147805   Mean   :127.2   Mean   :128.8   Mean   :29.28
##  3rd Qu.:156338   3rd Qu.:135.2   3rd Qu.:137.0   3rd Qu.:34.28
##  Max.   :159968   Max.   :159.0   Max.   :172.9   Max.   :69.81
##       wmed          age_min         age_max         age_med
```

```
##  Min.   :15.45   Min.   :3.000   Min.   : 8.03   Min.   : 5.185
##  1st Qu.:22.20   1st Qu.:3.060   1st Qu.:10.85   1st Qu.: 6.939
##  Median :26.60   Median :3.130   Median :13.22   Median : 8.463
##  Mean   :28.40   Mean   :3.166   Mean   :13.56   Mean   : 8.766
##  3rd Qu.:32.31   3rd Qu.:3.220   3rd Qu.:16.14   3rd Qu.:10.286
##  Max.   :69.40   Max.   :3.970   Max.   :19.99   Max.   :15.250
##       vnum           age_diff
##  Min.   : 11.00   Min.   : 5.00
##  1st Qu.: 34.00   1st Qu.: 7.68
##  Median : 47.00   Median :10.06
##  Mean   : 50.42   Mean   :10.40
##  3rd Qu.: 61.00   3rd Qu.:13.00
##  Max.   :187.00   Max.   :16.93
```

```
filter(epic_ind1)
```

```
## # A tibble: 1,272 x 10
##        id hmean  hmed wmean  wmed age_min age_max age_med  vnum age_diff
##     <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>   <dbl>   <dbl> <int>    <dbl>
##  1 103104  140.  144.  36.8  34.0    3.38    19.9    11.6    62     16.5
##  2 103125  137.  144   33.6  35.8    3.03    20.0    14.9    99     16.9
##  3 103145  143.  155   39.2  43.3    3.1     19.8    13.1    72     16.8
##  4 103148  145.  147.  44.4  44.8    3.02    18.2    11.2    62     15.2
##  5 103151  133.  133.  34.5  32.2    3.35    16.0    10.0    46     12.6
##  6 103187  150.  154.  42.2  39.8    3.08    19.8    11.5    55     16.7
##  7 103196  147.  153   44.7  44.7    3.15    19.1    13.2    55     15.9
##  8 103257  159.  168.  42.8  42.6    3.04    19.8    14.5   113     16.7
##  9 103258  144.  155   49.5  57.6    3.15    19.9    11.8    75     16.7
## 10 103290  143.  160   39.8  49.5    3.29    19.4    13.2    35     16.1
## # ... with 1,262 more rows
```

```
write.csv(epic_hw1, file = "data/epic/registration_age_min_3_4.csv")
```