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() %>%
  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
  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)
  # Tue Jan 04 09:34:32 2022 ------------------------------
  ## only select the visit number over 10 times
  ## 1325 individuals left so far

# nrow(epic_ind)
  # 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
```

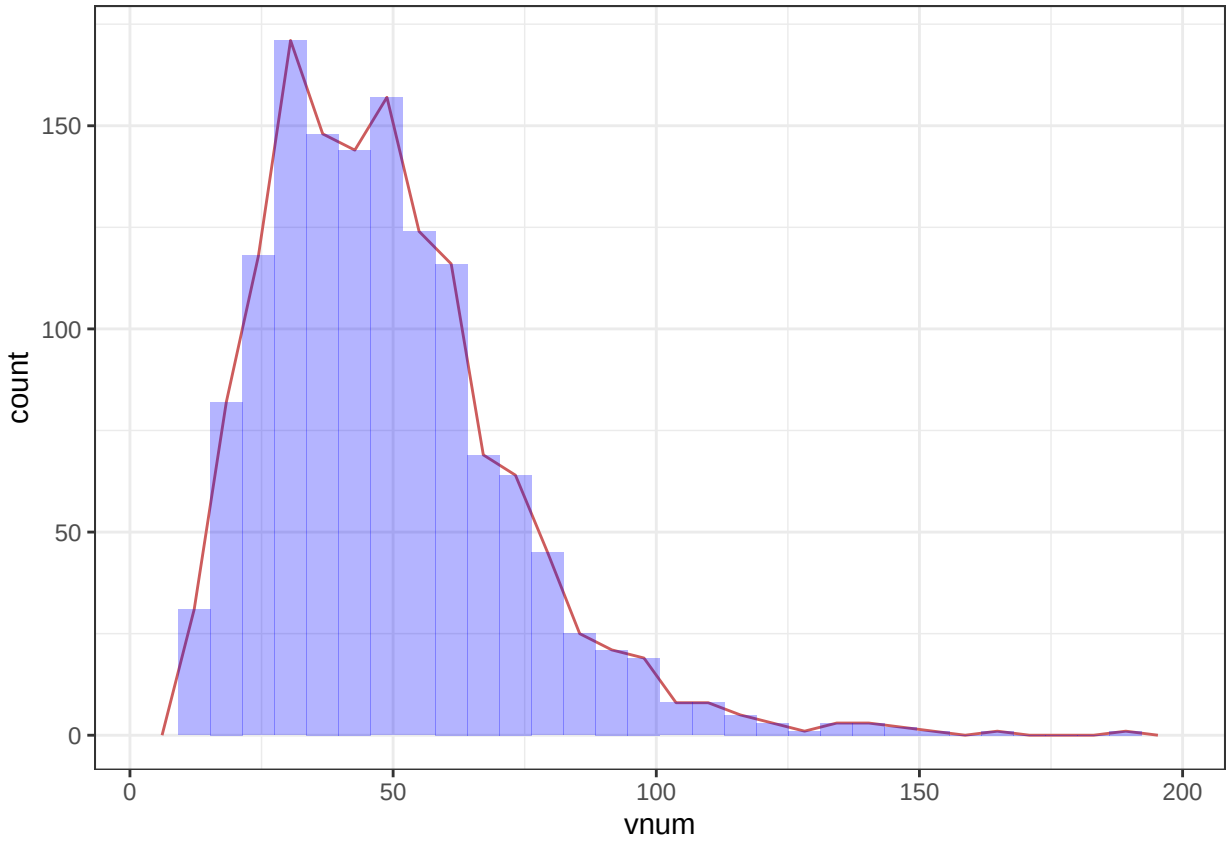
```
epic_ind1 %>% filter(age_diff >= 8) %>% nrow() ## 919
```

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

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

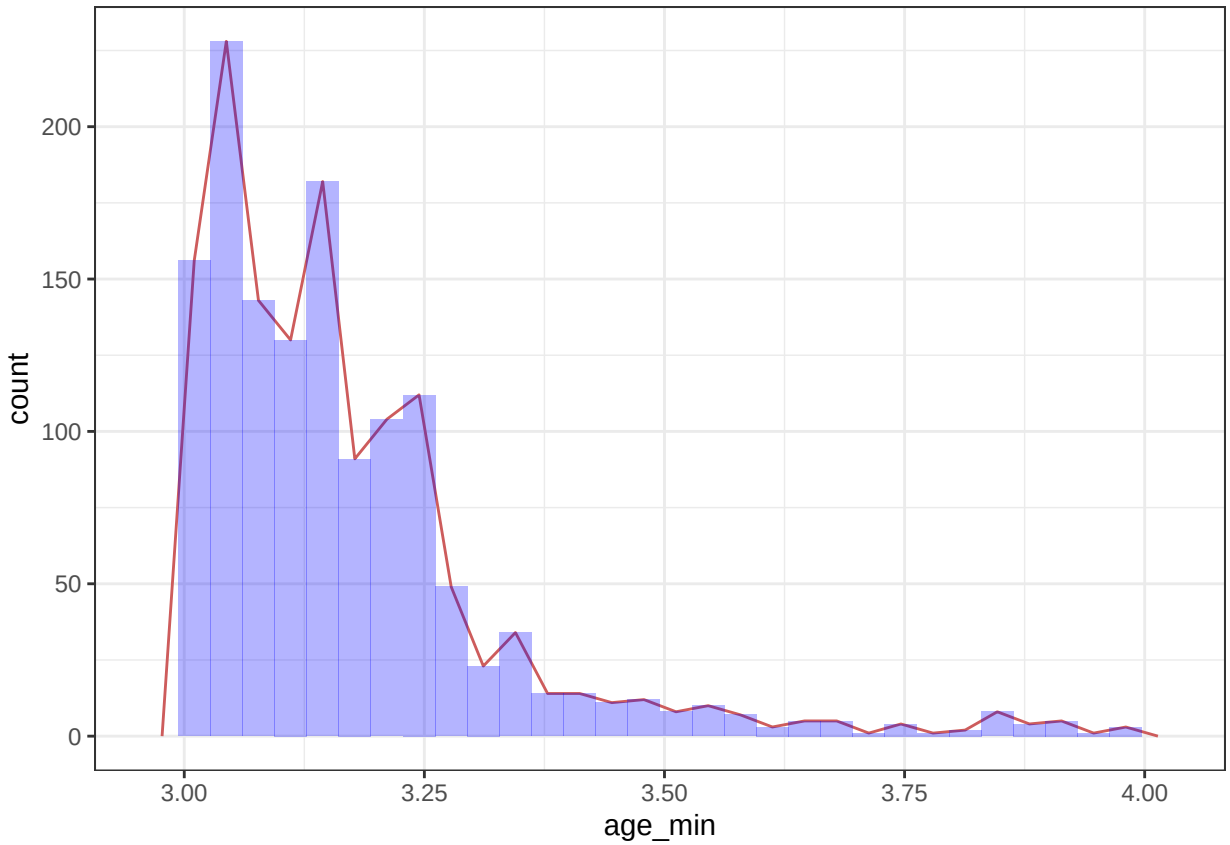
```
# View(epic_10_id)
# summary(epic_ind)

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`.
```



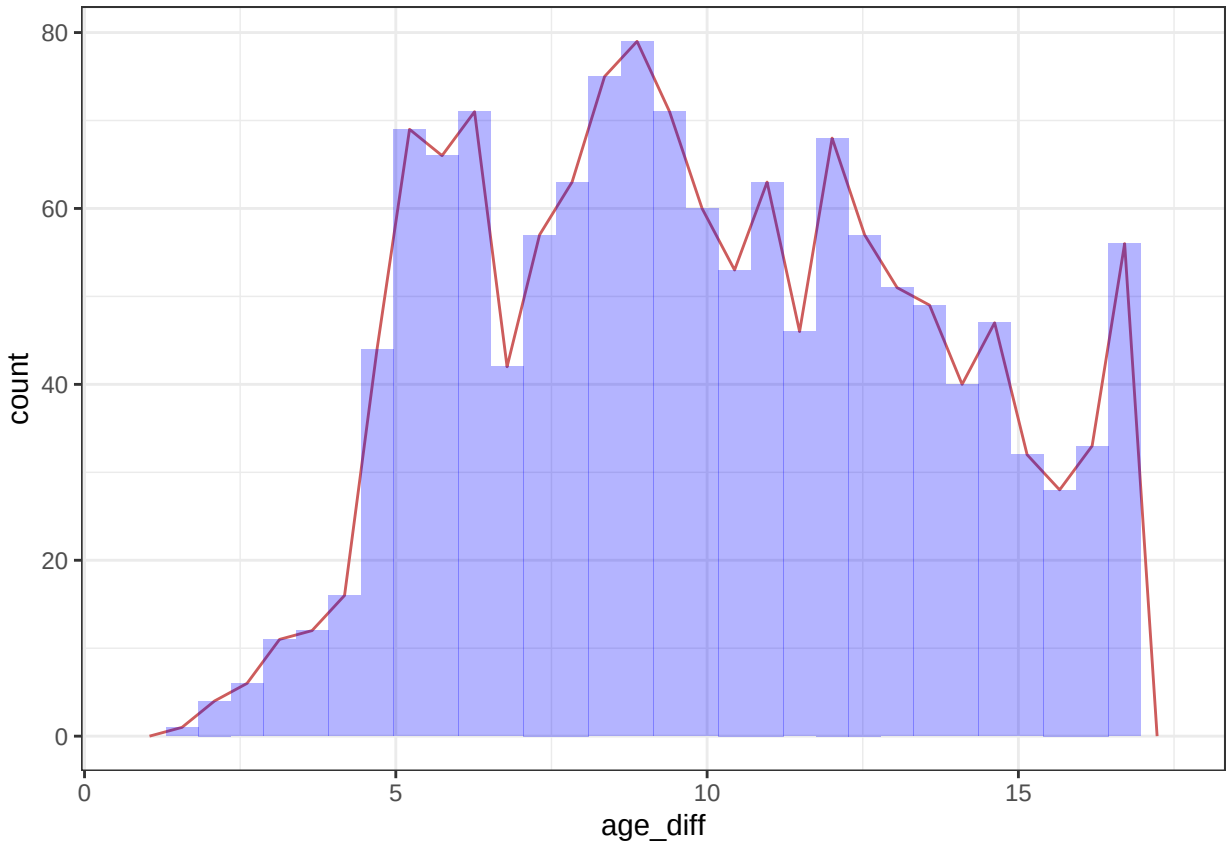

```
ggplot(data = epic_ind1, aes(age_min)) +
  geom_freqpoly(color = "indianred") +
  geom_histogram(fill = "blue", alpha = 0.3) +
  theme_bw()
```

```
## `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] 1370
```

```
summary(epic_ind1)
```

```
##        id             hmean            hmed            wmean
##  Min.   :103104   Min.   : 96.01   Min.   : 94.45   Min.   :12.83
##  1st Qu.:142795   1st Qu.:116.63   1st Qu.:117.21   1st Qu.:22.30
##  Median :152128   Median :125.31   Median :126.00   Median :26.97
##  Mean   :147745   Mean   :125.92   Mean   :127.43   Mean   :28.60
##  3rd Qu.:156234   3rd Qu.:134.47   3rd Qu.:136.00   3rd Qu.:33.68
##  Max.   :159968   Max.   :158.95   Max.   :172.90   Max.   :69.81
##       wmed          age_min          age_max         age_med
```

```
##  Min.   :12.50   Min.   :3.000   Min.   : 5.07   Min.   : 3.790
##  1st Qu.:21.45   1st Qu.:3.060   1st Qu.:10.23   1st Qu.: 6.631
##  Median :25.90   Median :3.130   Median :12.79   Median : 8.178
##  Mean   :27.75   Mean   :3.168   Mean   :13.11   Mean   : 8.514
##  3rd Qu.:31.84   3rd Qu.:3.220   3rd Qu.:15.98   3rd Qu.:10.094
##  Max.   :69.40   Max.   :3.970   Max.   :19.99   Max.   :15.250
##       vnum           age_diff
##  Min.   : 10.00   Min.   : 1.790
##  1st Qu.: 32.00   1st Qu.: 7.050
##  Median : 45.00   Median : 9.635
##  Mean   : 48.31   Mean   : 9.945
##  3rd Qu.: 60.00   3rd Qu.:12.720
##  Max.   :187.00   Max.   :16.930
```

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