HW4

Songyu Tang

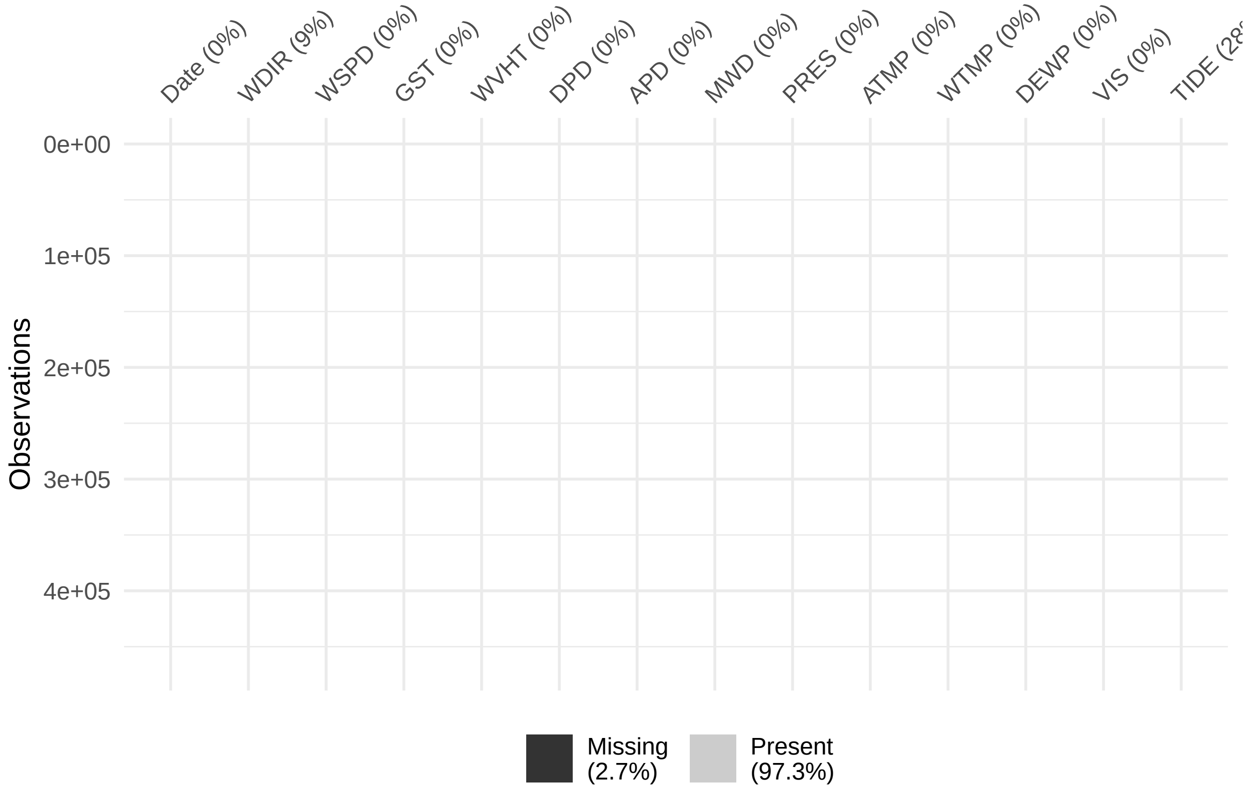
Fall 2024

(b)

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(naniar)
buoy_data <- read.csv("buoy_data_1985_2023.csv")
buoy_data <- buoy_data %>%
  mutate(WDIR = ifelse(WDIR == 999, NA, WDIR))
#No, it is not appropriate to convert missing/null data to NA. In this data, if 999 represents an outli
vis_miss(buoy_data,warn_large_data = FALSE)
```

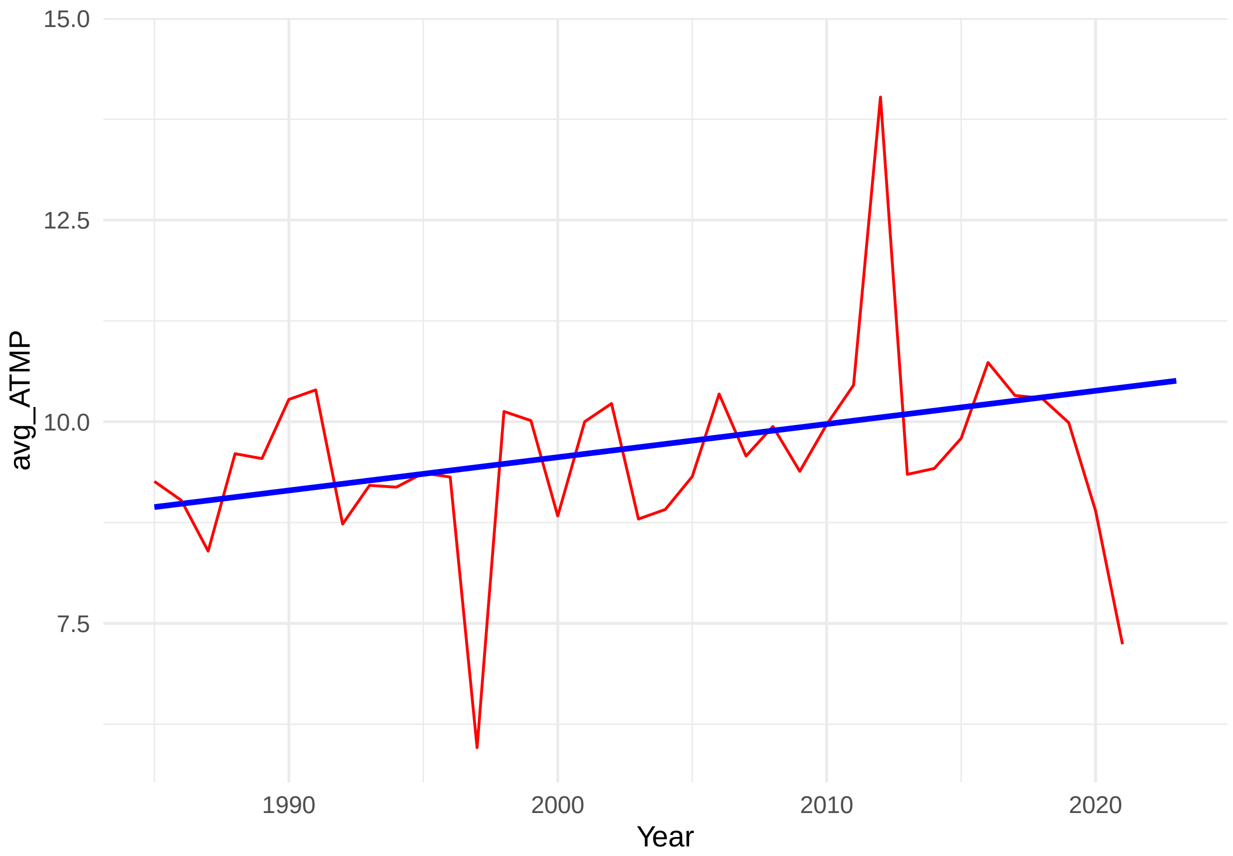


(c)

```r
#I think plot ATMP and WTMP may be a good choice to show the effect of climate change.
library(tidyverse)
buoy_data <- read.csv("buoy_data_1985_2023.csv")
buoy_data <- buoy_data %>%
  mutate(Year = year(Date))
climate_summary <- buoy_data %>%
  mutate(ATMP = ifelse(ATMP == 999, NA, ATMP)) %>%
  mutate(WTMP = ifelse(WTMP == 999, NA, WTMP)) %>%
  group_by(Year) %>%
  summarize(
    avg_ATMP = mean(ATMP, na.rm = TRUE),
    avg_WTMP = mean(WTMP, na.rm = TRUE)
  )
ggplot(climate_summary, aes(x = Year, y = avg_ATMP)) +
  geom_line(col = "red") +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  theme_minimal()
```

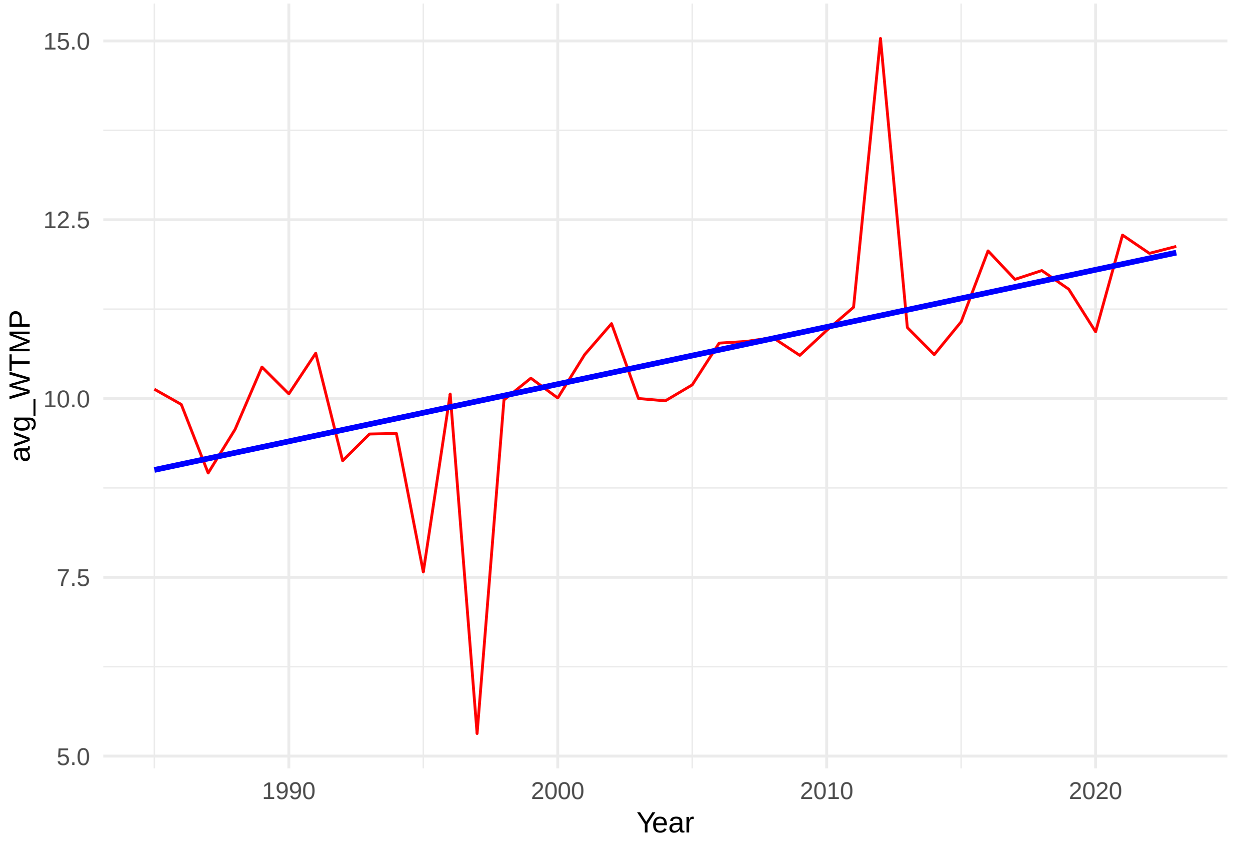
```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_smooth()`).
```



```r
ggplot(climate_summary, aes(x = Year, y = avg_WTMP)) +
  geom_line(col = "red") +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



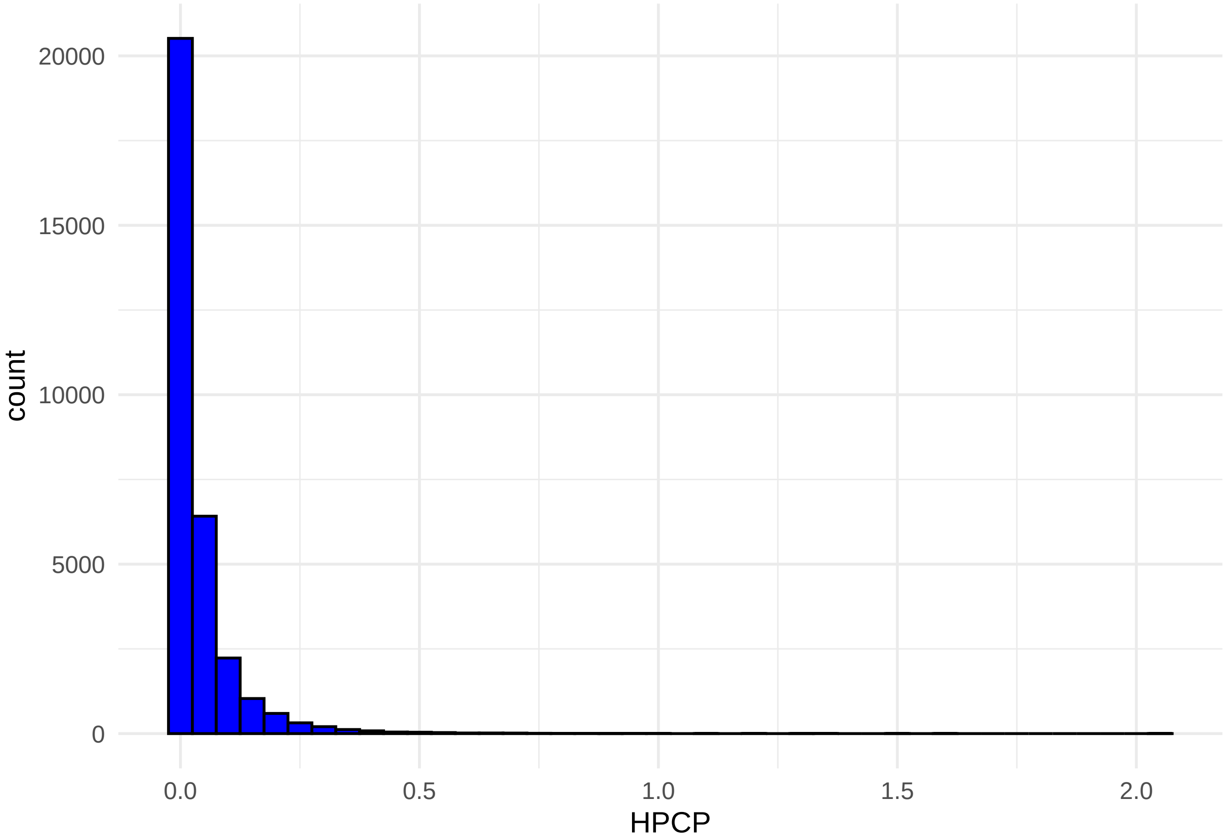
```
#From the plot, it seems that both ATMP and WTMP have positive slopes in linear regression. It can conc
```

(d)

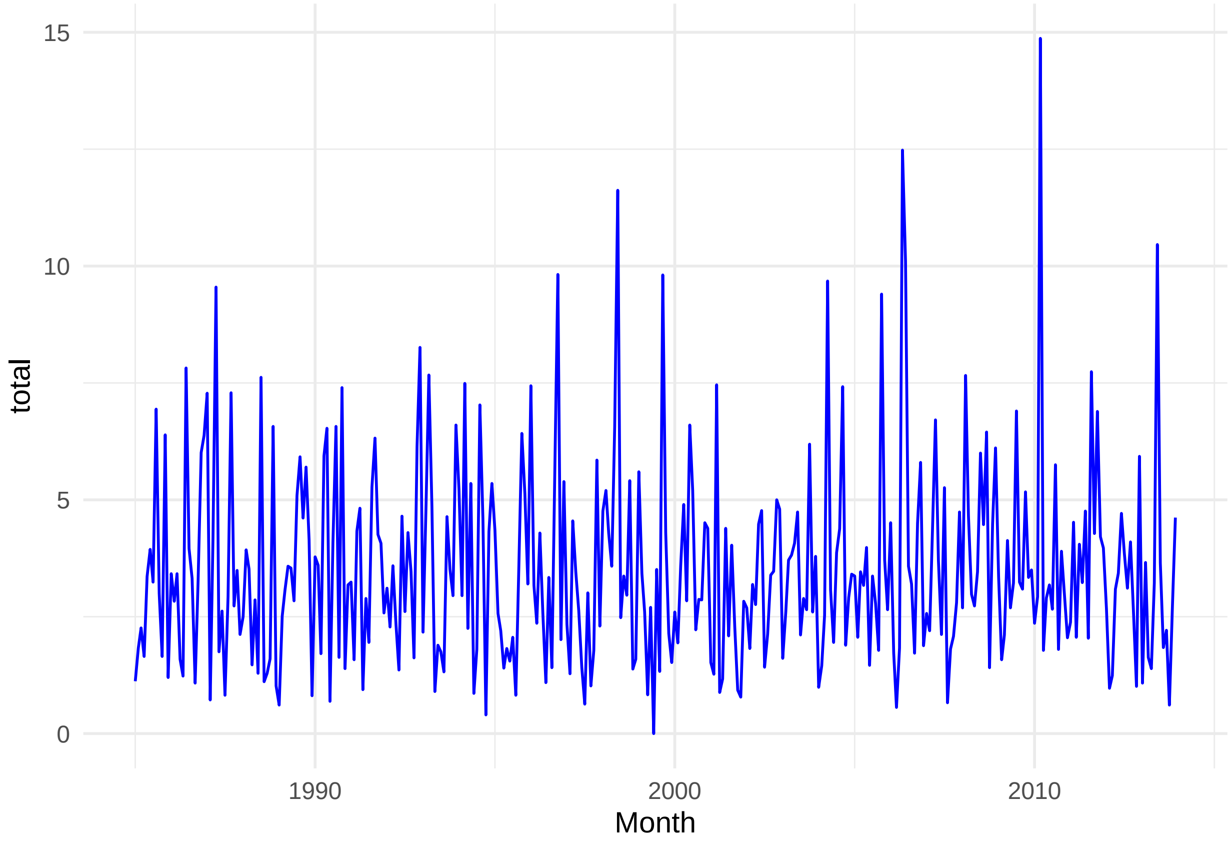
```r
library(tidyverse)
rainfall <- read.csv("Rainfall.csv")
rainfall <- rainfall %>%
  mutate(DATE = ymd_hm(DATE))
rainfall_cleaned <- rainfall %>%
  select(DATE, HPCP)
summary(rainfall_cleaned$HPCP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.01000 0.03875 0.04000 2.03000
```

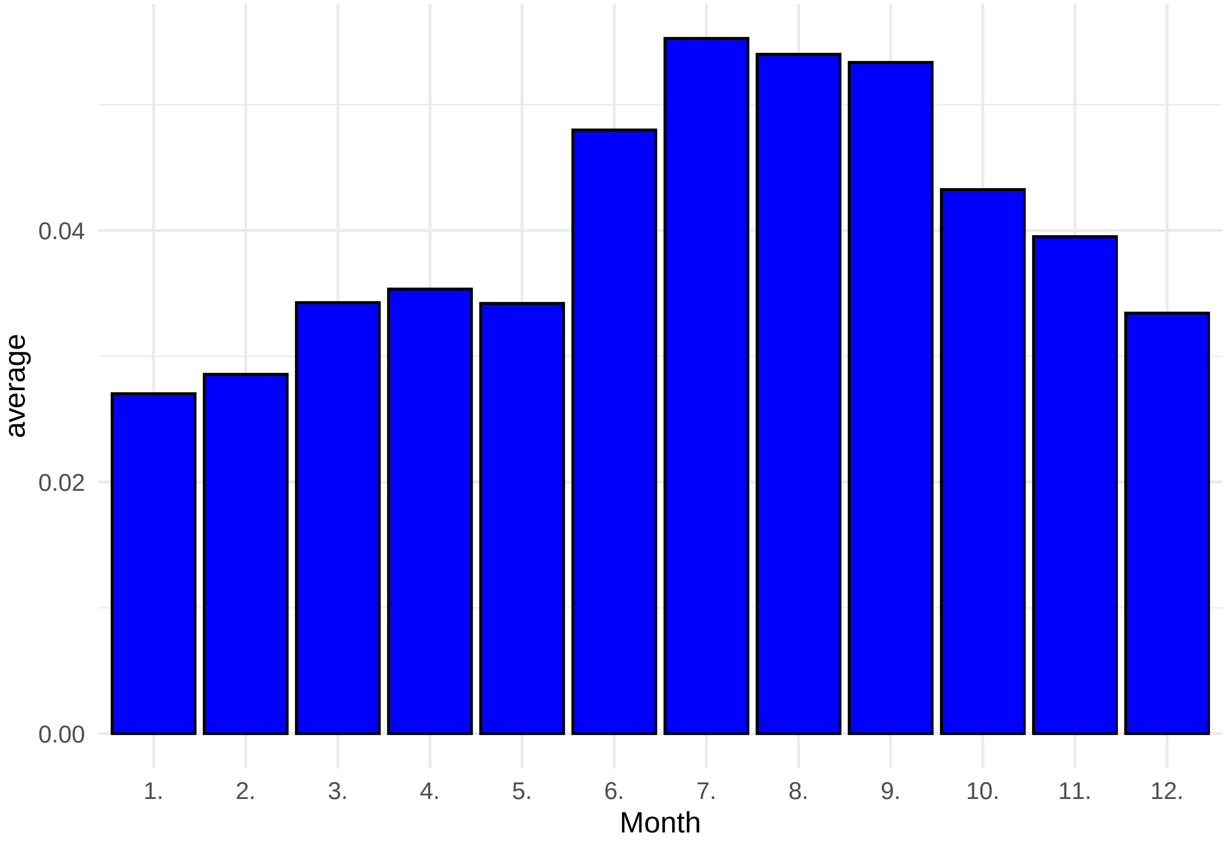
```r
#plot the histogram of the count of HPCP
ggplot(rainfall_cleaned, aes(x = HPCP)) +
  geom_histogram(binwidth = 0.05, fill = 'blue', color = 'black') +
  theme_minimal()
```



```r
rainfall_month <- rainfall_cleaned %>%
  mutate(Month = floor_date(DATE,"month")) %>%
  group_by(Month) %>%
  summarise(total = sum(HPCP, na.rm = TRUE))
#plot the change in the HPCP corresponding to the month
ggplot(rainfall_month, aes(x = Month, y = total)) +
  geom_line(color = 'blue') +
  theme_minimal()
```



```
rainfall_month_average <- rainfall_cleaned %>%
  mutate(Month = month(DATE, label = TRUE)) %>%
  group_by(Month) %>%
  summarise(average = mean(HPCP, na.rm = TRUE))
#plot the total HPCP in different month
ggplot(rainfall_month_average, aes(x = Month, y = average)) +
  geom_bar(stat = 'identity', fill = 'blue', color = 'black') +
  theme_minimal()
```



```r
#assume the HPCP has relationship with date, so make a simple regression about date and HPCP
rainfall_model <- lm(HPCP ~ as.numeric(DATE), data = rainfall_cleaned)
summary(rainfall_model)
```

```
##
## Call:
## lm(formula = HPCP ~ as.numeric(DATE), data = rainfall_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.05756 -0.03536 -0.02598  0.00284  1.97310
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.471e-02  1.753e-03   42.62   <2e-16 ***
## as.numeric(DATE) -3.622e-11  1.713e-12  -21.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07582 on 31712 degrees of freedom
## Multiple R-squared:  0.01391,    Adjusted R-squared:  0.01387
## F-statistic: 447.2 on 1 and 31712 DF,  p-value: < 2.2e-16
```