Ceiling

Adam Stuller

CIG SKY-CONDITION-OBSERVATION ceiling height dimension

Výška polohy oblakov v metroch nad zemou.

```
all_data <- read.csv(file= "../data/all.csv")
```

```
describe(all_data$CIG)
```

```
## all_data$CIG
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##   292566   120770      217    0.977     7360     9113       61      210
##      .25      .50      .75      .90      .95
##      750     1707    22000    22000    22000
##
## lowest :     0    15    25    30    60, highest:  7200  7500  9000 19500 22000
```

Centralna poloha dat

Hodnota vyberoveho medianu je 1707, modus je 22000 a vyberovy priemer je 7359.877. Tieto tri hodnoty hovoria o tom, ze data urcite nie su vyvazene ani sa nezhromazduju okolo silneho stredu. Vydime, ze najcastejsia hodnota je 22000, čo je ďaleko vyššie ako median. Priemer je zjavne vychýlený smerom doprava.

Spôsobené je to tým, že hodnoty zaznamenane sa postupne zvysuju podla toho ako su namerane. Od isteho bodu si vsat vsetky hodnoty 22000. To moze znamenat, ze oblaky uz boli vyssie ako ich senzor bol schopny zaznamenat alebo, ze obloha bola uplne jasna.

```
getmode(na.omit(all_data$CIG)) %>%
  print(cat("Modus: " ))
```

```
## Modus: [1] 22000
```

```
median(all_data$CIG, na.rm = TRUE) %>%
  print(cat("Median: "))
```

```
## Median: [1] 1707
```

```
mean(all_data$CIG, na.rm = TRUE) %>%
  print(cat("Mean: "))
```

```
## Mean: [1] 7359.877
```

Variabilita

Variacne rozpatie je 22 000 teda od 0 ked predpokladame, ze je hmla po jasnu oblohu.

Medzi kvartilova odchylka je 10625, co znamena, ze 50 percent dat je velmi siroko rozdelenych.

Vyberovy rozptyl je 87070999, štandardna odchylka 9331.184 a variancny koeficient 1.267845, cize 126.78%, čo je naozaj veľka variabilita.

```r
max_slp <- max(all_data$CIG, na.rm= TRUE)
min_slp <- min(all_data$CIG, na.rm= TRUE)
var_rozpatie <- max_slp - min_slp
print(cat("Variacne rozpatie", var_rozpatie))
```

```
## Variacne rozpatie 22000NULL
```

```r
# Interquartile range
Q1_slp <- quantile(all_data$CIG, 0.25, na.rm = T) # 25% hodnot je mensich a 75% vacsich
Q3_slp <- quantile(all_data$CIG, 0.75, na.rm = T) # 75% hodnot je mensich a 25% vacsich

(IQR(all_data$CIG, na.rm = T ) / 2) %>%# interquartile range
  print(cat("Medzikvantilova odchýlka: "))
```

```
## Medzikvantilova odchýlka: [1] 10625
```

```r
var(all_data$CIG, na.rm = T) %>% print(cat("Rozptyl: "))# rozptyl
```

```
## Rozptyl: [1] 87070999
```

```r
EnvStats::cv(all_data$CIG, na.rm = T) %>% print(cat("Variacny koeficient: "))# variacny koeficient
```

```
## Variacny koeficient: [1] 1.267845
```

```r
summary(all_data$CIG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0     750    1707    7360   22000   22000  120770
```

```r
all_data['CIG'] %>% profiling_num()
```

```
##   variable     mean  std_dev variation_coef p_01 p_05 p_25 p_50  p_75  p_95
## 1      CIG 7359.877 9331.184       1.267845   30   61  750 1707 22000 22000
##    p_99  skewness kurtosis   iqr    range_98     range_80
## 1 22000 0.8968205 1.869349 21250 [30, 22000] [210, 22000]
```

Asymetria

Šikmost (skewness) je 0.8968205. Je teda kladna a hovori, ze data su nachylene dolava.

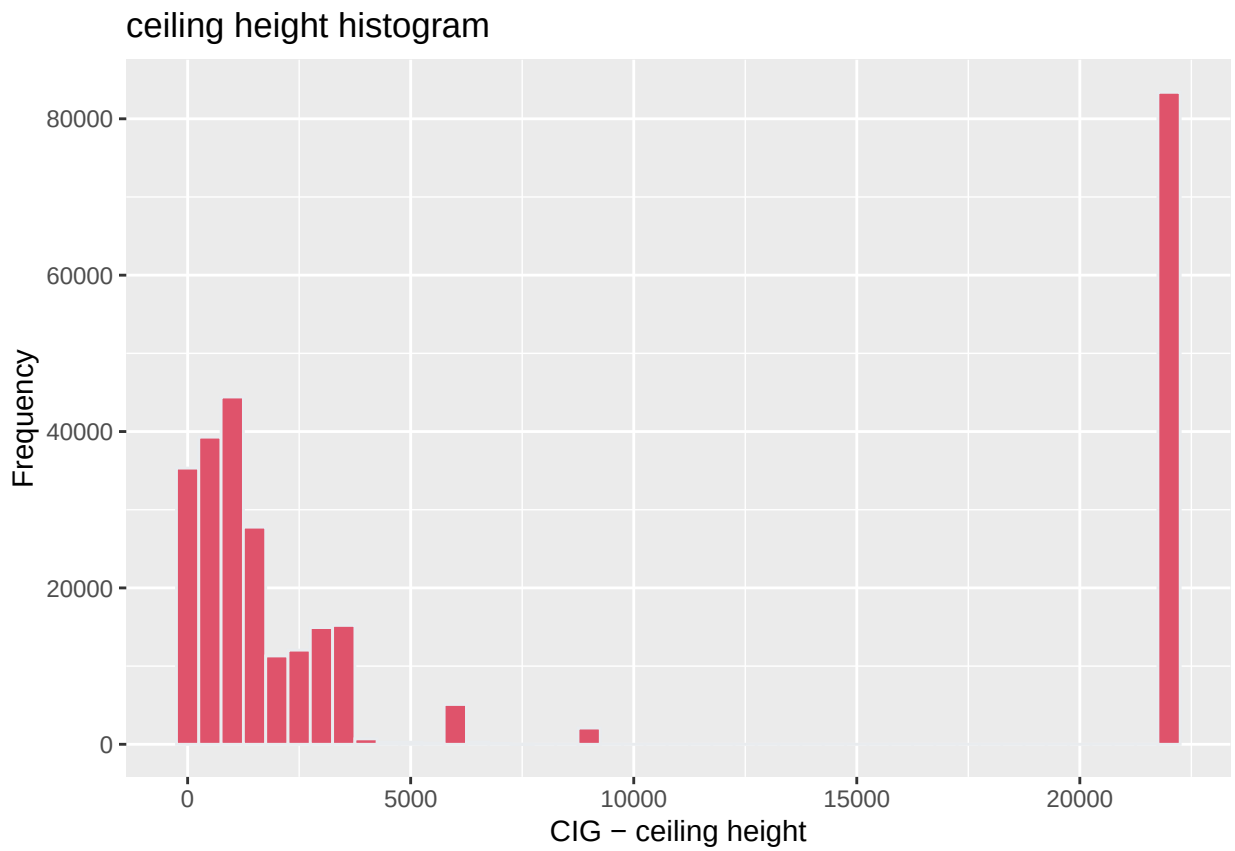
Špicatost (kurtosis) - 1.869349. AKo ale ukazuje histogram, CIG ma velmi daleko od normalneho rozdelenia a tieto veliciny nam vela nepovedia.

Histogram

Vidime, ze distribucia dat je velmi zaujimava.

```r
ggplot(all_data, aes(x=CIG)) +
  geom_histogram(bins = 10, binwidth = 500,fill="2", color="#e9ecef") +
  labs(title = paste("ceiling height histogram")) +
  xlab("CIG - ceiling height") +
  ylab("Frequency")
```

```
## Warning: Removed 120770 rows containing non-finite values (stat_bin).
```



Pri pohlade na hodnoty nizsie ako maximalna hodnota 22000, vidime este lepsie, ze sa to nepotoba na normalne ani na ine slusne rozdelnie.

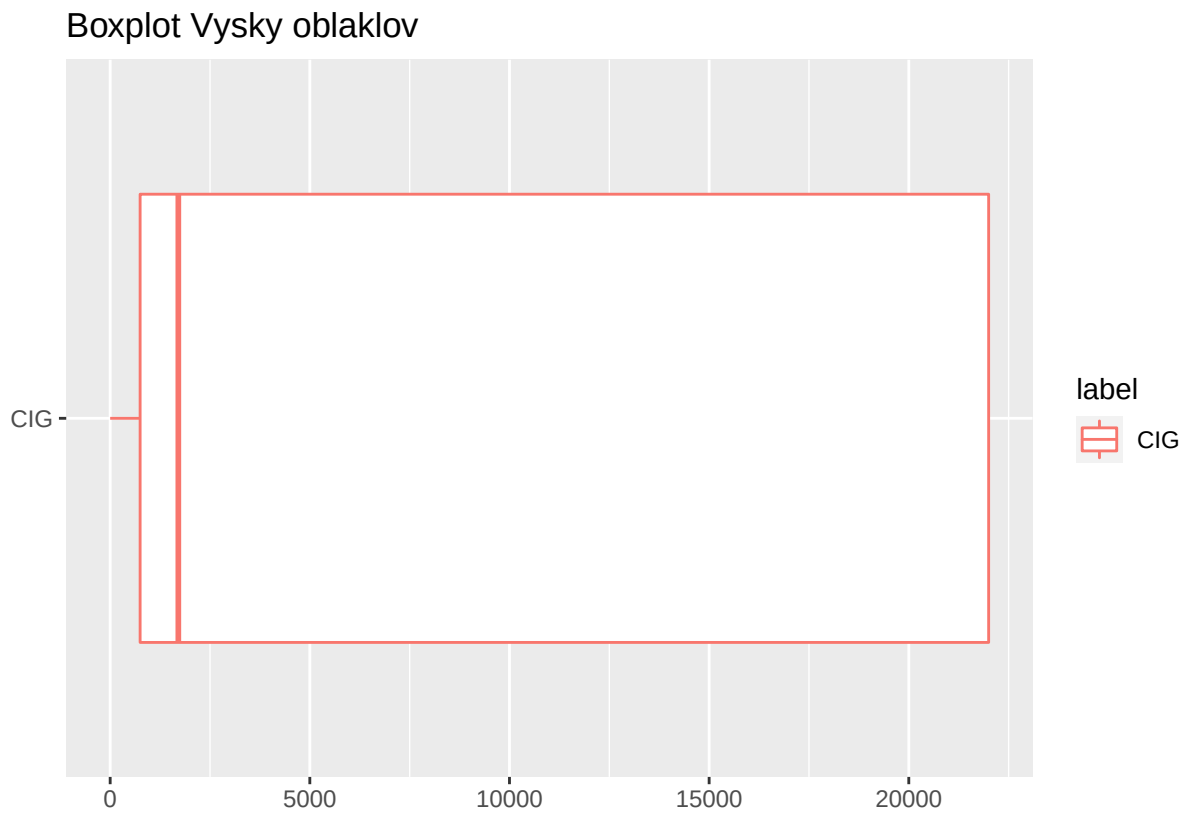
Boxplot

Pro pohlade na boxplot vidime, ze data su vysoko nevyvazene a median je daleko od stredu grafu.

```
df <- all_data %>%
  dplyr::select('CIG') %>%
  tidyr::gather(key='label', value = 'ceiling')

ggplot(data = df, aes( ceiling,factor(label), colour=label)) +
  geom_boxplot() +
  labs(title = paste("Boxplot Vysky oblaklov")) +
  xlab("") +
  ylab("")
```

```
## Warning: Removed 120770 rows containing non-finite values (stat_boxplot).
```

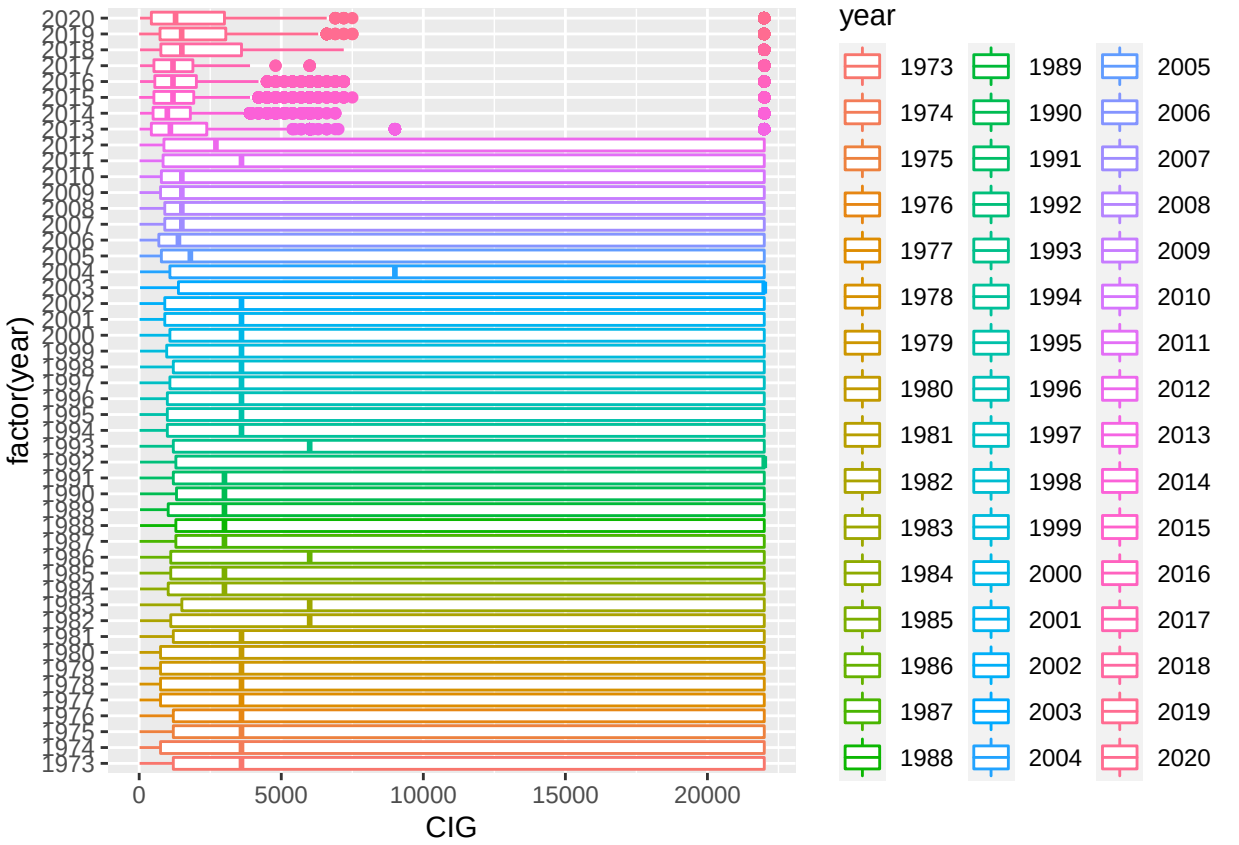


Na boxplote pre jednotlive roky vidime, ze po roku 2012 prislo k zaujimavej zmene merania.

```
df <- all_data %>%
  dplyr::mutate(
    year = ymd_hms(DATE) %>%
      lubridate::year() %>%
      map_chr(~ as.character(.x))
  ) %>%
  dplyr::select(all_of(c('year', 'CIG')))

ggplot(data = df, aes( CIG,factor(year), colour=year)) +
  geom_boxplot()
```

```
## Warning: Removed 120770 rows containing non-finite values (stat_boxplot).
```



Violin

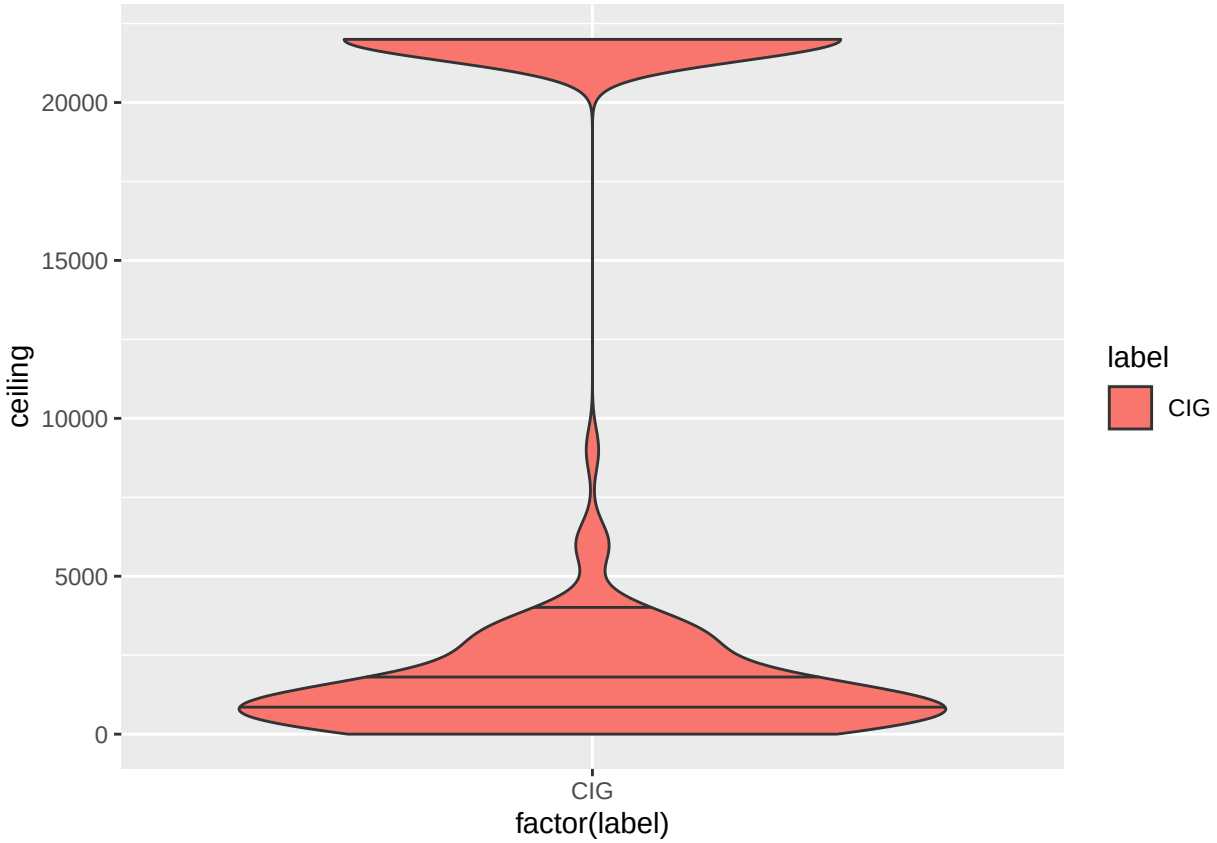
Violin plot ukazuje ze okrem hodnoty 22000 sa merania zvyknu zhromazdovat este niekde pod 1000 a nad 5000.

```
df <- all_data %>%
  dplyr::select('CIG') %>%
  tidyr::gather(key='label', value = 'ceiling')

ggplot(data = df, aes(factor(label), ceiling, fill=label)) +
  geom_violin(draw_quantiles=c(0.25, 0.5, 0.75))
```

```
## Warning: Removed 120770 rows containing non-finite values (stat_ydensity).
```

```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```



Q-Q plot

qq plot iba ukazuje, ze o normalnom rozdelni nemoze byt ani rec.

```
ggplot(data = all_data, aes(sample=CIG)) +
  stat_qq() +
  stat_qq_line()
```

```
## Warning: Removed 120770 rows containing non-finite values (stat_qq).
```

```
## Warning: Removed 120770 rows containing non-finite values (stat_qq_line).
```

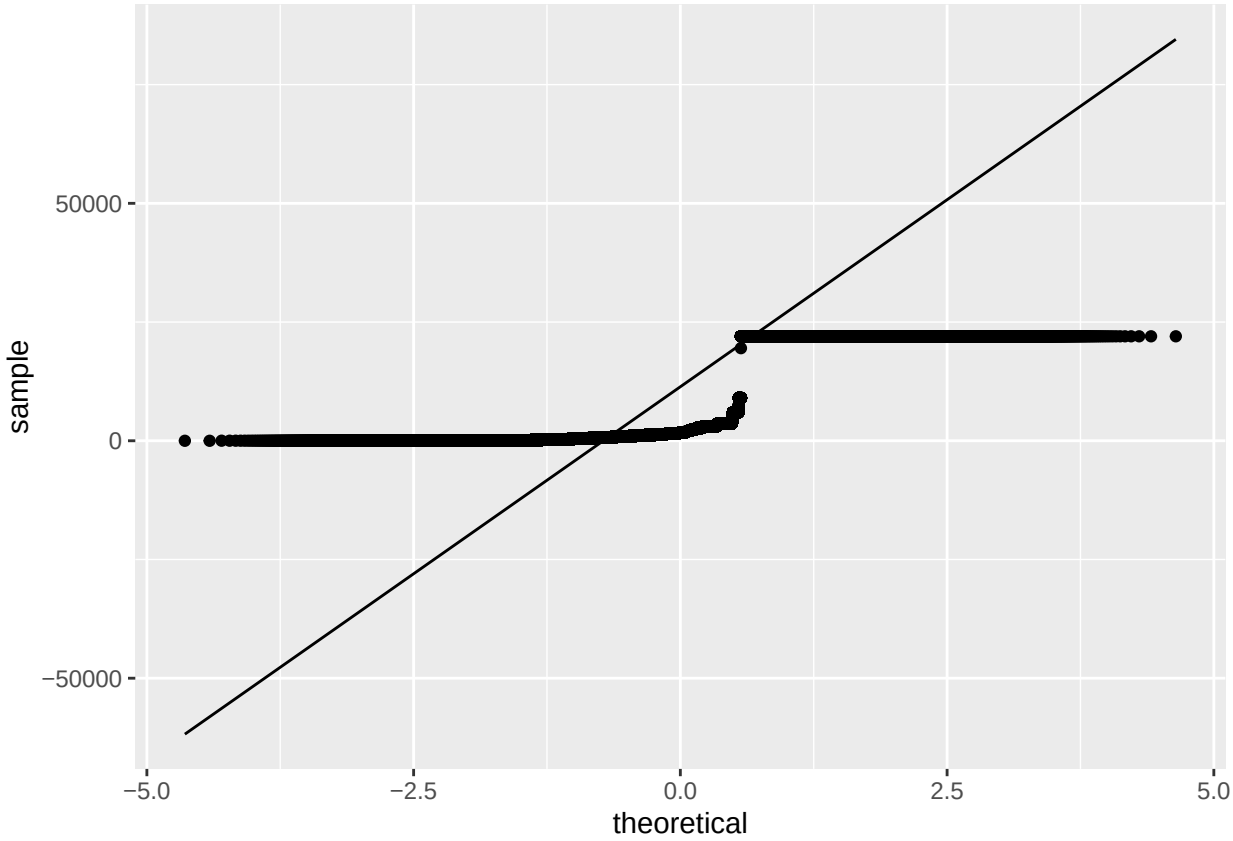


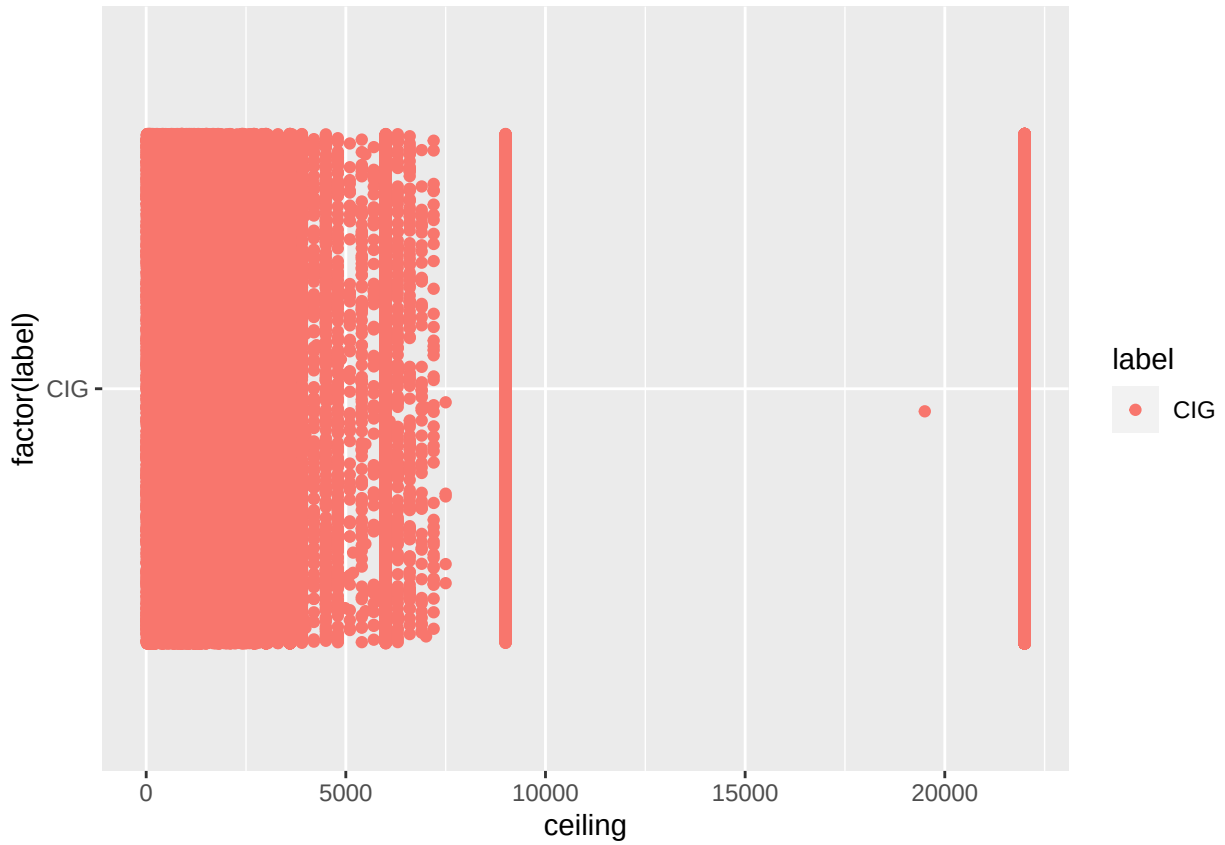
Diagram Rozptylenia

Ukazuje sa nam aj druhe časte meranie pod 10000, ktoré sme uz vyššie spomenuli. Nevieme čoho je to nasledkom.

```
df <- all_data %>%
  dplyr::select('CIG') %>%
  tidyr::gather(key='label', value = 'ceiling')

ggplot(data = df, aes( ceiling,factor(label), colour=label)) +
  geom_jitter()
```

```
## Warning: Removed 120770 rows containing missing values (geom_point).
```

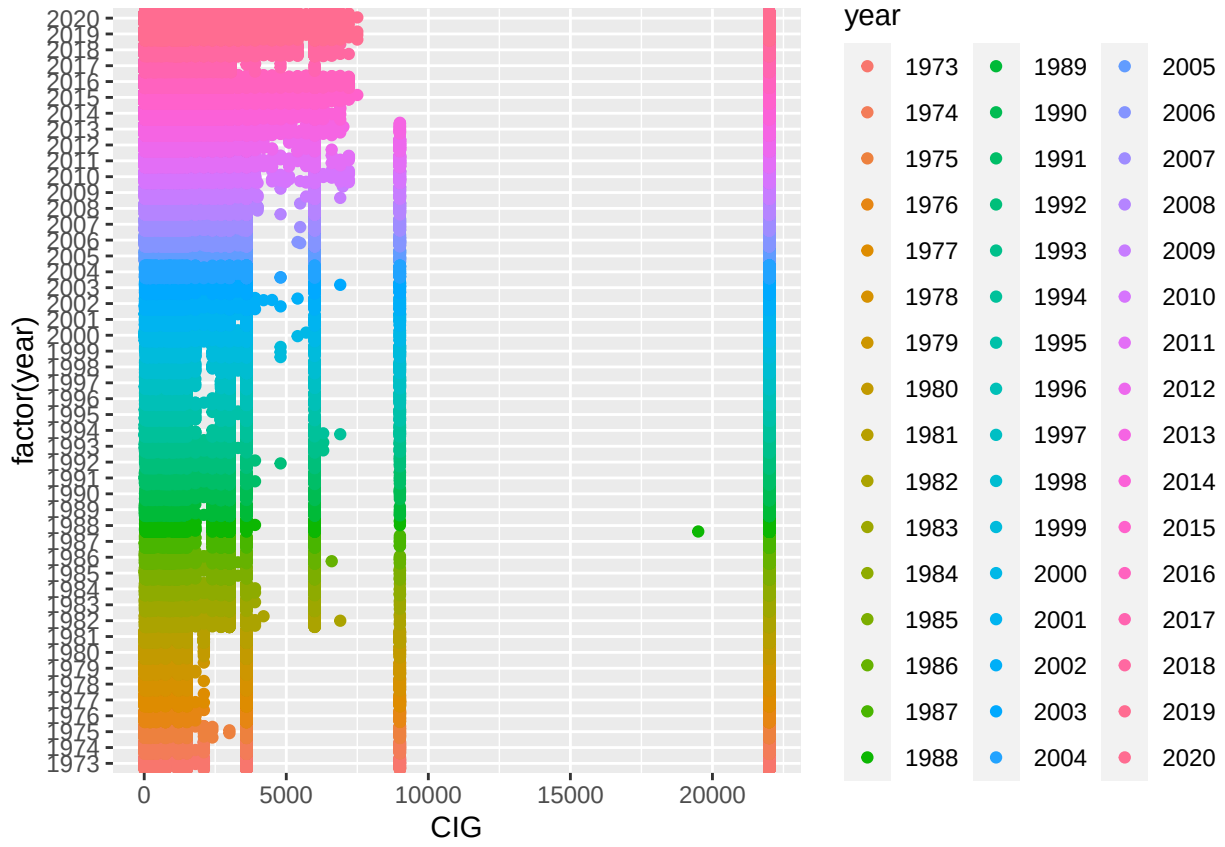


Vidime vsak na diagrame rozptylenia pre jednotlive roky, ze sa to dialo pred uz spominanym rokom 2013.

```
df <- all_data %>%
  dplyr::mutate(
    year = ymd_hms(DATE) %>%
      lubridate::year() %>%
      map_chr(~ as.character(.x))
  ) %>%
  dplyr::select(all_of(c('year', 'CIG')))

ggplot(data = df, aes( CIG,factor(year), colour=year)) +
  geom_jitter()
```

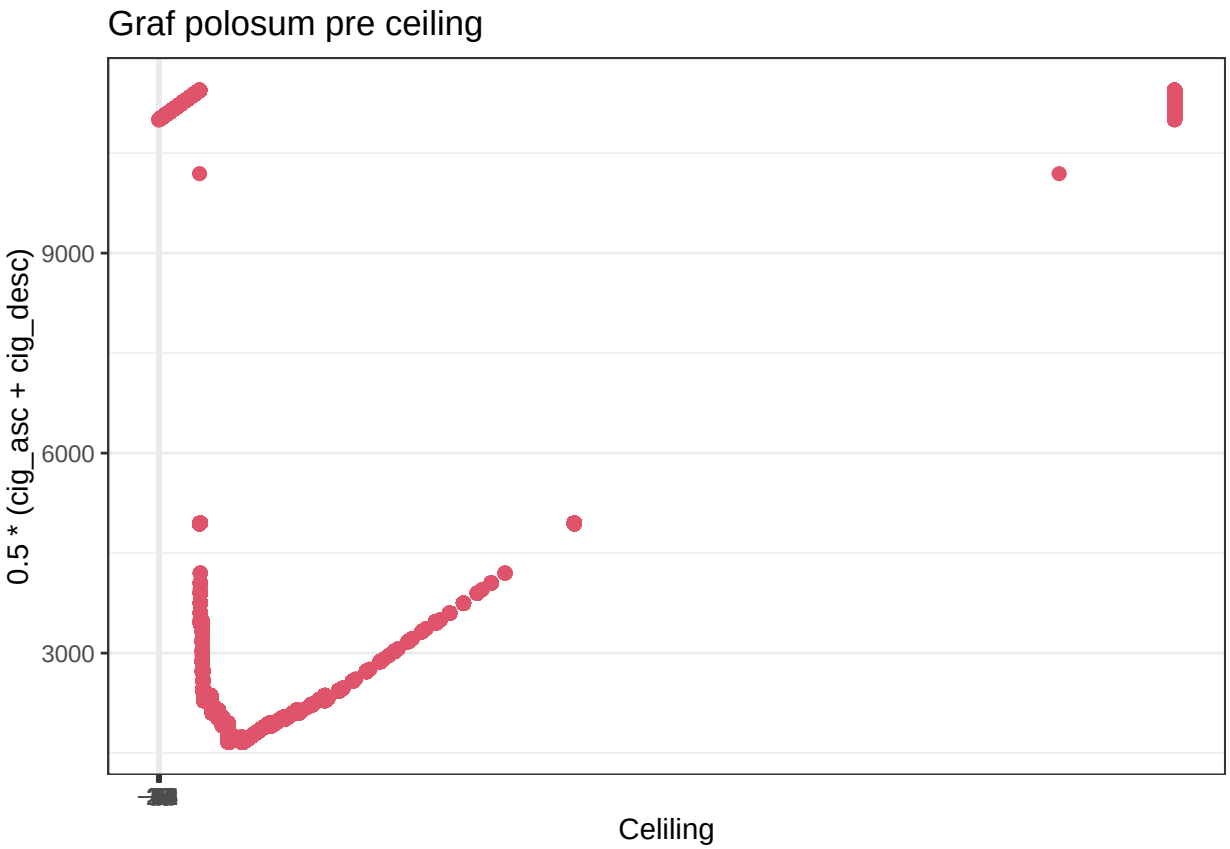
```
## Warning: Removed 120770 rows containing missing values (geom_point).
```



Graf polosum

```
cig <- all_data$CIG
cig_asc <- sort(cig, decreasing = FALSE)
cig_desc <- sort(cig, decreasing = TRUE)

ggplot(data.frame(cig_asc), aes(x = cig_asc, y = 0.5*(cig_asc+cig_desc))) +
  geom_point(size = 2, color = 2) +
  scale_x_continuous(breaks = seq(-30, 40, by = 3)) +
  labs(title = "Graf polosum pre ceiling", x = "Celiling") +
  theme_bw()
```


Graf polosum pre ceiling
0.5 * (cig_asc + cig_desc)
9000
6000
3000
Celiling