Visibility

Denisa Mensatorisova

VISIBILITY-OBSERVATION distance dimension

The horizontal distance at which an object can be seen and identified.

MIN: 000000 MAX: 160000 UNITS: Meters

Missing = 999999

NOTE: Values greater than 160000 are entered as 160000

Viditeľnosť predstavuje horizontálnu vzdialenosť, po ktorú je objekt viditeľný meranú v metroch.

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

```
summary(all_data$VIS)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0    6000   10000   12841   15000   75000   89684
```

```
all_data['VIS'] %>% profiling_num()
```

```
##   variable     mean  std_dev variation_coef p_01 p_05 p_25  p_50  p_75  p_95
## 1      VIS 12840.67 9716.109      0.7566665  100  600 6000 10000 15000 30000
##    p_99 skewness kurtosis  iqr     range_98      range_80
## 1 40000 1.313705 4.979899 9000 [100, 40000] [2000, 25000]
```

```
getmode(na.omit(all_data$VIS)) # modus
```

```
## [1] 9900
```

Centrálna poloha dát

Hodnota výberového mediánu je 10000 a hodnota výberového priemeru 12841 Keďže stredná hodnota (medián) je nižšia ako priemer, hodnotu priemeru mohli ovplyvniť vyššie hodnoty.

Modus - najčastejšia hodnota viditeľnosti je 9900.

```
var(all_data$VIS, na.rm = T) # rozptyl
```

```
## [1] 94402766
```

```
max(all_data$VIS, na.rm = T) - min(all_data$VIS, na.rm = T) # variacne rozpatie
```

```
## [1] 75000
```

```
# Interquartile range and outliers
Q1 <- quantile(all_data$VIS, 0.25, na.rm = T) # 25% hodnot je mensich a 75% vacsich
Q3 <- quantile(all_data$VIS, 0.75, na.rm = T) # 75% hodnot je mensich a 25% vacsich
IQR <- IQR(all_data$VIS, na.rm = T) # interquartile range
IQR_dev <- IQR/2
```

```r
# odlahle hodnoty
length(which(all_data$VIS < (Q1 - 1.5*IQR)))
```

```
## [1] 0
```

```r
length(which(all_data$VIS > (Q3 + 1.5*IQR)))
```

```
## [1] 31632
```

```r
# extremne hodnoty
length(which(all_data$VIS < (Q1 - 3*IQR)))
```

```
## [1] 0
```

```r
length(which(all_data$VIS > (Q3 + 3*IQR)))
```

```
## [1] 3176
```

Variabilita

Výberový rozptyl je 94402766.

Výberová smerodajná odchýlka je 9716.109. To znamená, že viditeľnosť sa pohybuje približne v rozsahu 9716.109 okolo priemeru.

Medzikvartilová odchýlka (IQR/2) je 4500, teda hodnoty sú rozptýlené približne 4500 okolo mediánu.

Variačné rozpätie je 75000. Daná hodnota predstavuje rozdiel medzi maximálnou a minimálnou nameranou hodnotou viditeľnosti. Maximálna viditeľnosť je 75000 a minimálna 0.

Variačný koeficient viditeľnosti je 0.76, čo je 76%.

Asymetria

Hodnota šikmosti je kladná 1.313705, to znamená, že viac hodnôt sa nachádza pod priemerom. Dáta sú zošimené a nejde o symetrické rozdelenie okolo strednej hodnoty.

Hodnota špicatosti je 4.9799, je väčšia ako 3 teda hodnoty majú špicaté rozdelenie. To znamená, že v súbore sa nachádza viac hodnôt bližšie k strednej hodnote.

Boxplot

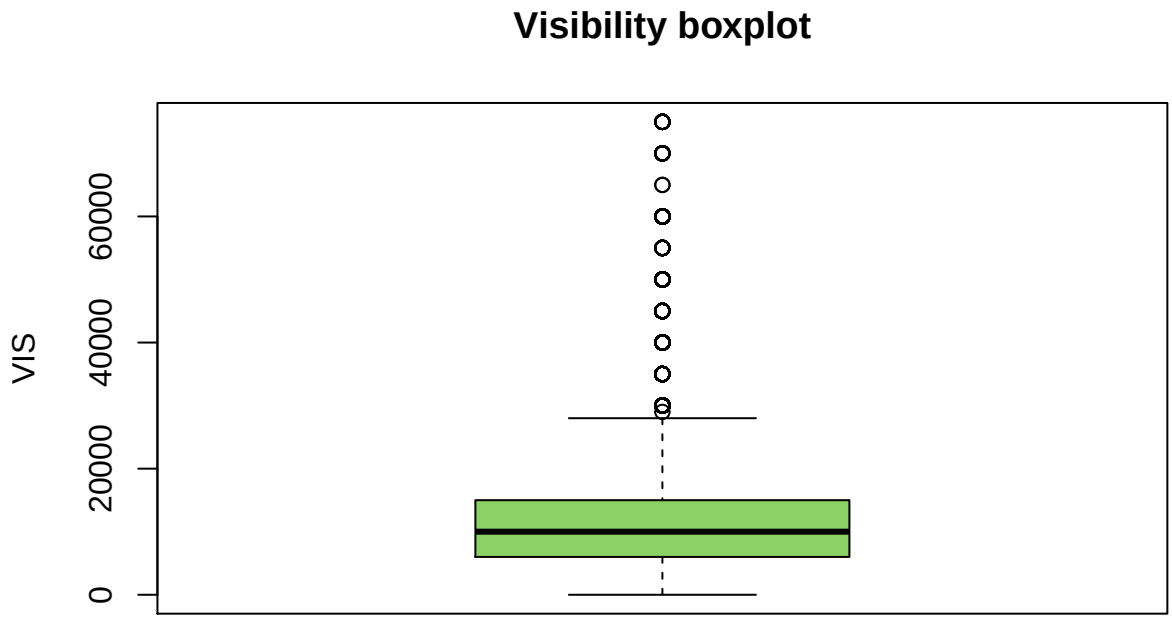
Graf rozdeľuje namerané hodnoty teploty na niekoľko častí. Hodnota 3.kvartilu je 15000 a hodnota 1. kvartilu 6000, teda medzikvartilové rozpätie (IQR) je 9000. Uprostred krabice je zvýraznený medián hrubou čiernou čiarou (9900). Kedže sa nachádza mierne pod polovicou krabice, podľa boxplotu sa zdá, že dáta sú mierne zošikmené. Tiež vidíme, že krabica boxplotu sa nachádza v nižších hodnotách, teda 50% hodnôt tvoria nízke hodnoty, avšak je tu aj niekoľko vysokých hodnôt.

Ďalej z boxplotu vidieť maximálnu a minimálnu hodnotu (vonkajšie hradby boxplotu). Maximálna hodnota (28500) je vypočítaná ako 3.kvartil + 1.5 * IQR (medzikvartilové rozpätie). Minimálna hodnota je vypočítaná ako 1.kvartil - 1.5 * IQR (medzikvartilové rozpätie), v tomto prípade je to 0.

Všetky hodnoty nachádzajúce sa nad a pod maximálnou a minimálnou hodnotou môžme považovať za odľahlé hodnoty. Počet odľahlých hodnôt nad maximálnou hodnotou je 31632, pod minimálnou sa nenachádzajú žiadne odľahlé hodnoty.

Nakoniec pre odľahlé hodnoty overíme či patria medzi extrémne. Horná hranica extrémych hodnôt je vypočítaná ako 3.kvartil + 3 * IQR. Dolná hranica extrémych hodnôt je vypočítaná ako 1.kvartil - 3 * IQR. V dátach sa nachádza 3176 extrémne vysokých hodnôt, ktoré sú vyššie ako hodnota 3.kvartilu + 3 * IQR, teda vyššie ako 42000.

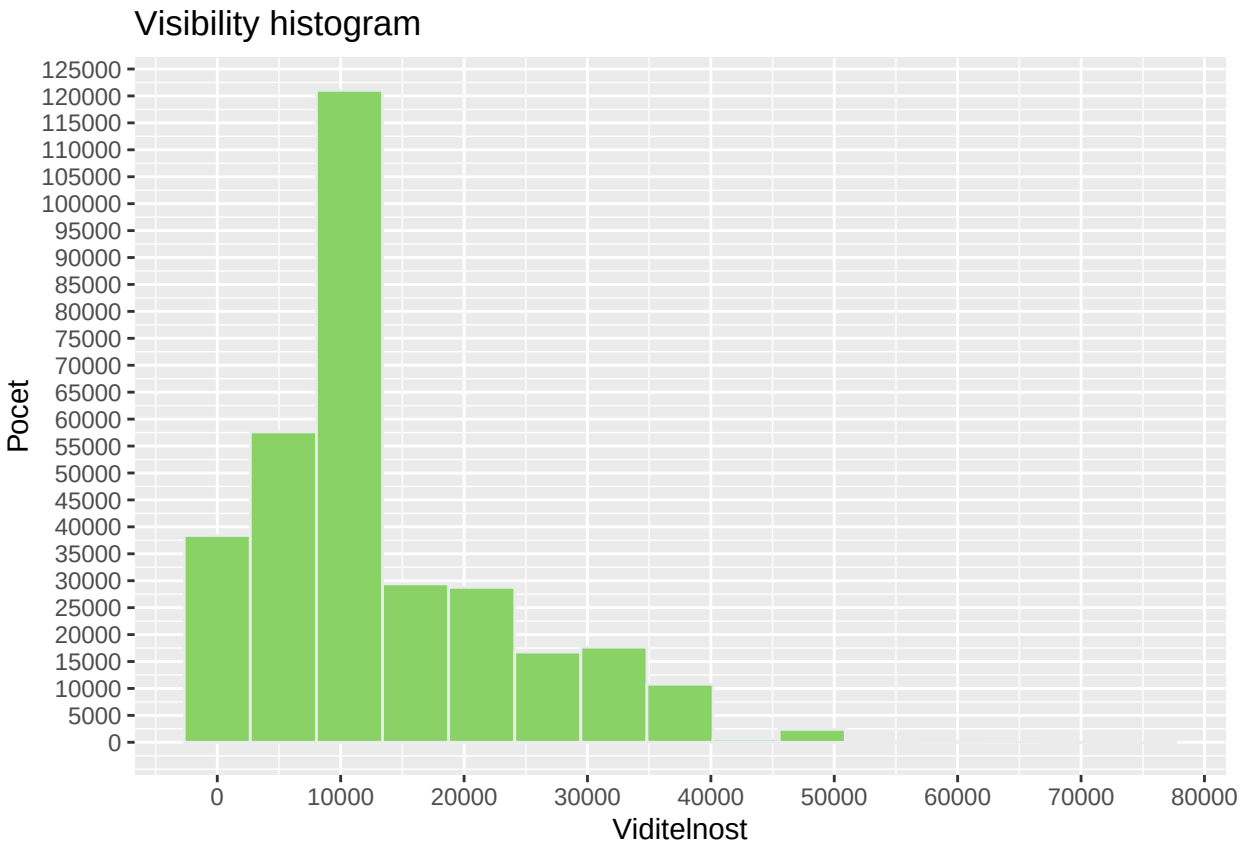
```r
boxplot(all_data$VIS, col = "#8ad166", ylab = "VIS", main = "Visibility boxplot")
```



Histogram

Aj z histogramu vidíme, že rozdelenie hodnôt je vyšikmené doľava. Najpočetnejšie sú hodnoty viditeľnosti okolo 10000 metrov.

```r
ggplot(all_data, aes(x = VIS)) +
  geom_histogram(bins = 15, fill = "#8ad166", color = "#e9ecef") +
  labs(title = paste("Visibility histogram")) +
  xlab("Viditeľnosť") +
  ylab("Počet") +
  scale_x_continuous(breaks = seq(0, 80000, by = 10000)) +
  scale_y_continuous(breaks = seq(0, 150000, by = 5000))
```

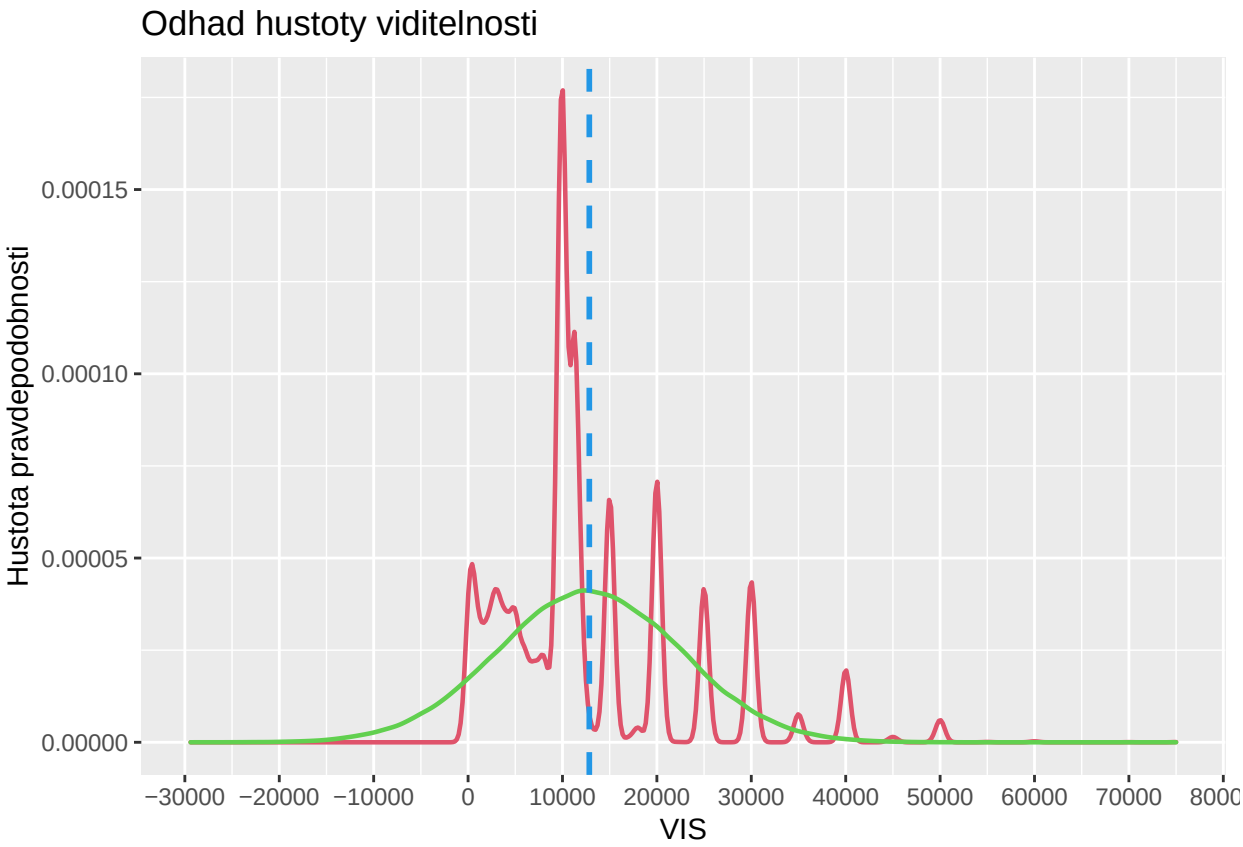


Graf hustoty

Graf hustoty slúži na porovnanie priebehu hustoty pravdepodobnosti normálneho rozdelenia (zelená čiara) a odhadu hustoty vypočítaného z namernaých hodnôt viditeľnosti (červená čiara). Čiary nie sú rovnaké, teda nejde o normálne rozdelenie. Modrá prerušovaná čiara predstavuje priemernú hodnotu viditeľosti. V dátach sa nachádza niekoľko vrcholov, najvyšší je okolo hodnoty 10000.

```
# denisty plot
# data z normalneho rozdelenia
data_norm <- data.frame(dens = c(rnorm(length(na.omit(all_data$VIS)), mean(all_data$VIS, na.rm = T), sd

# porovnanie hodnot normalneho rozdelenia a VIS
ggplot(all_data, aes(x = VIS), color = 3) +
  geom_density(color = 2, size = 0.8) +
  geom_density(data_norm, mapping = aes(x = dens), color = 3, size = 0.8) +
  geom_vline(aes(xintercept = mean(VIS, na.rm = T)),
             color = 4, linetype = "dashed", size = 1) +
  scale_x_continuous(breaks = seq(-60000, 80000, by = 10000)) +
  labs(title = paste("Odhad hustoty viditeľnosti")) +
  xlab("VIS") +
  ylab("Hustota pravdepodobnosti")
```



Husľový graf

Husľový graf doplnený o hlavné kvartily zobrazuje rozdelenie hustoty, pričom aj podľa tohto grafu vidíme, že nejde o normálne rozdelenie. Dáta sú najpočetnejšie v okolí hodnoty 10000, postupne smerom k vyšším hodnotám sa ich hustota zmenšuje.

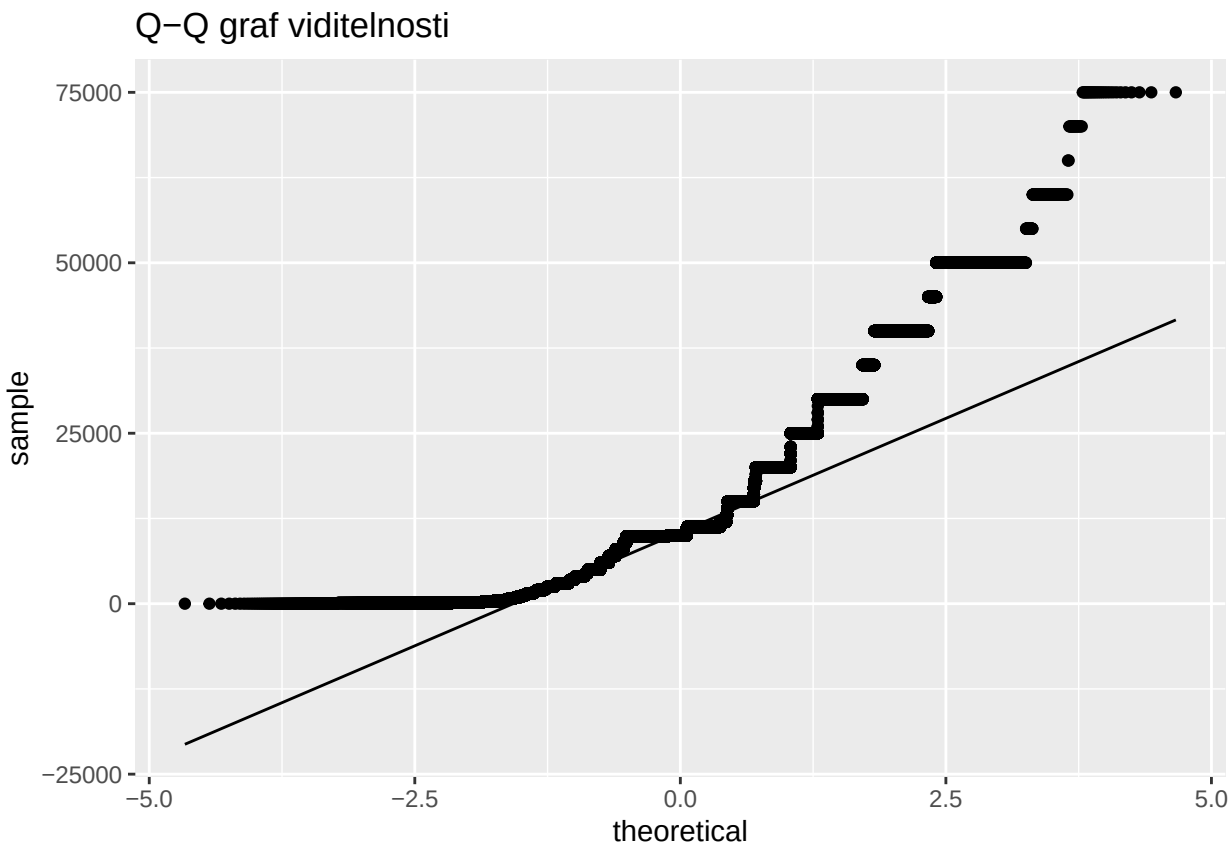
```
df <- all_data %>%
  dplyr::select('VIS') %>%
  tidyr::gather(key = 'label', value = 'vis')

ggplot(data = df, aes(factor(label), vis, fill = vis)) +
  geom_violin(draw_quantiles = c(0.25, 0.5, 0.75), fill = "#8ad166") +
  labs(title = paste("Husľový graf viditeľnosti"), y = "visibility", fill = "visibility") +
  theme(axis.title.x = element_blank()) +
  theme(axis.text.x = element_blank()) +
  scale_y_continuous(breaks = seq(0, 80000, by = 10000))
```

Q-Q graf

Graf zobrazuje odchýlku empirického od teoretického normálneho rozdelenia. Empirické rozdelenie je v našom prípade rozdelenie nameraných hodnôt viditeľnosti. Keďže body sa najmä v okolí 0 odchyľujú od priamky normálneho rozdelenia, môžeme povedať, že rozdelenie hodnôt viditeľnosti nie je normálne.

```
ggplot(data = all_data, aes(sample = VIS)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = paste("Q-Q graf viditeľnosti"))
```

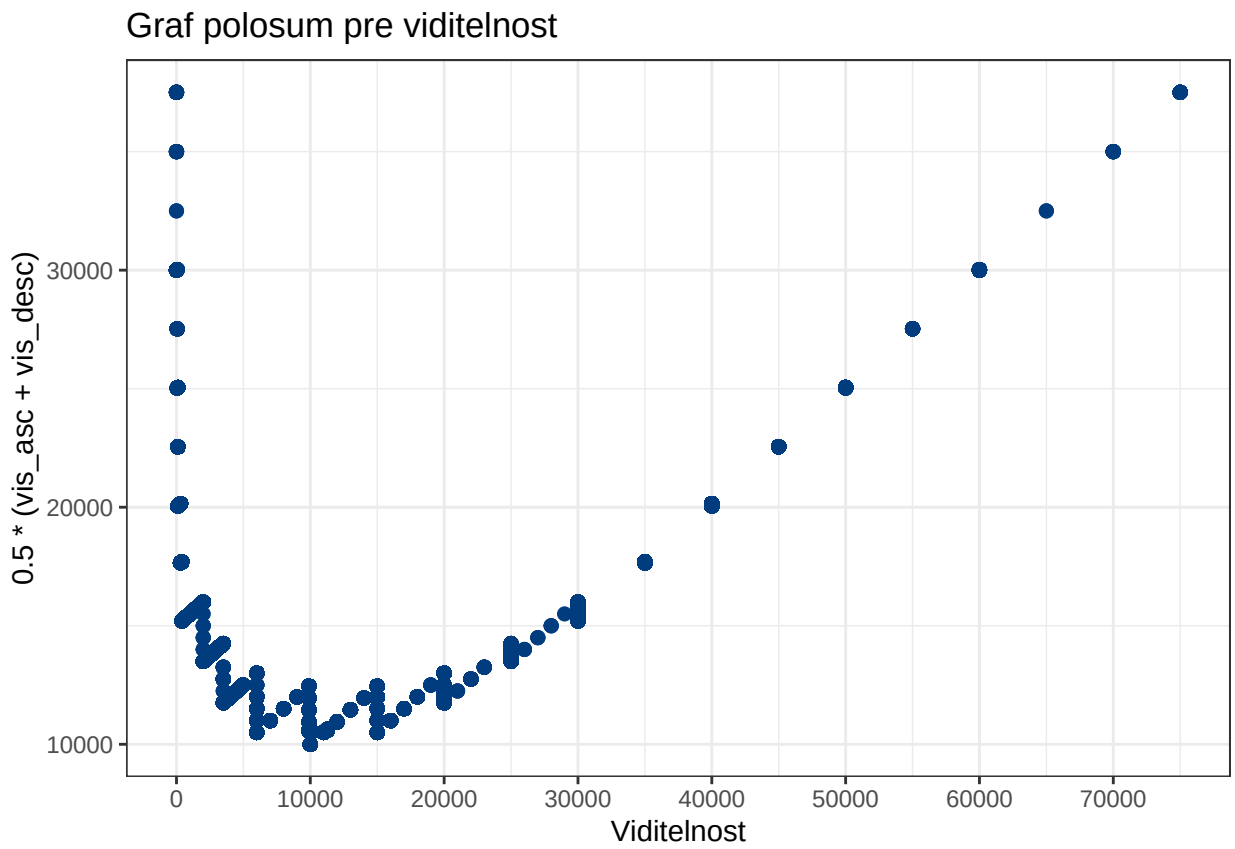


Graf polosum

Podobne aj z grafu polosum vidno, že hodnoty nie sú symetrické.

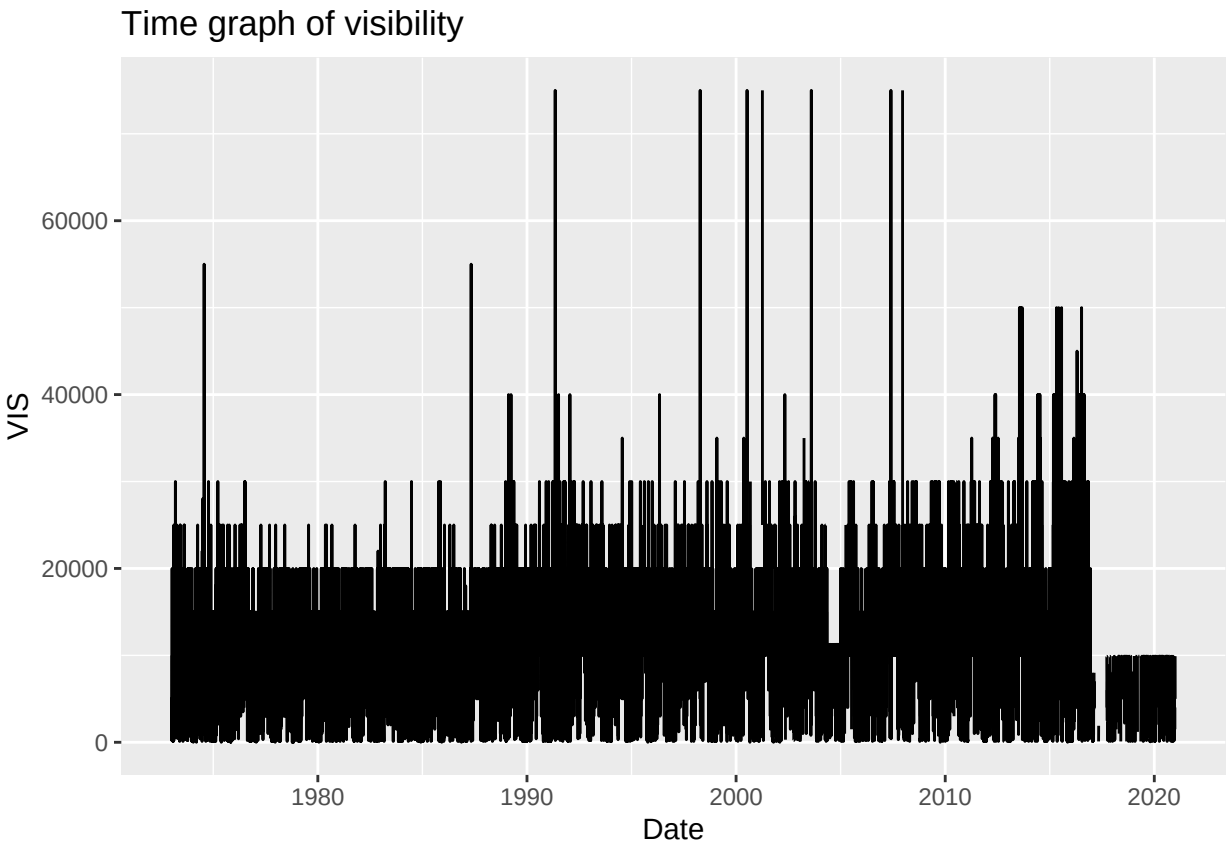
```
vis <- all_data$VIS
vis_asc <- sort(vis, decreasing = FALSE)
vis_desc <- sort(vis, decreasing = TRUE)

ggplot(data.frame(vis_asc), aes(x = vis_asc, y = 0.5*(vis_asc + vis_desc))) +
  geom_point(size = 2, color = "#013c7f") +
  scale_x_continuous(breaks = seq(0, 80000, by = 10000)) +
  labs(title = "Graf polosum pre viditeľnosť", x = "Viditeľnosť") +
  theme_bw()
```



Časový graf viditeľnosti

```
all_data %>%
  dplyr::mutate(
    date = as_date(DATE)
  ) %>%
  dplyr::distinct(date, .keep_all=TRUE) %>%
  dplyr::select(date, VIS) %>%
  as_tsibble(
    index = date
  ) %>%
  autoplot(VIS) +
  labs(title = "Time graph of visibility",
       y = "VIS", x = "Date")
```


Time graph of visibility
VIS
60000
40000
20000
0
1980
1990
2000
2010
2020
Date