

Sea level pressure

Denisa Mensatorisova

ATMOSPHERIC-PRESSURE-OBSERVATION sea level pressure

The air pressure relative to Mean Sea Level (MSL).

MIN: 08600 MAX: 10900 UNITS: Hectopascals

SCALING FACTOR: 10

99999 = Missing

Atribút sea level pressure reprezentuje atmosférický tlak meraný v hektopascaloch.

```
all_data <- read.csv(file= "../data/all.csv")  
  
describe(all_data$SLP)  
  
## all_data$SLP  
##      n    missing  distinct     Info      Mean      Gmd      .05      .10  
##  188602    224734      686        1    1017    9.146    1004    1008  
##      .25      .50      .75      .90      .95  
##      1012     1017     1022     1028     1032  
##  
## lowest :  970.6  971.4  971.9  972.6  973.4, highest: 1048.5 1048.6 1048.7 1049.3 1049.4  
summary(all_data$SLP)  
  
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.    NA's  
##  970.6 1012.0 1016.9 1017.3 1022.2 1049.4 224734  
all_data['SLP'] %>% profiling_num()  
  
##      variable      mean    std_dev variation_coef    p_01    p_05   p_25    p_50    p_75  
## 1      SLP 1017.295 8.221255 0.008081483 997.5 1004.5 1012 1016.9 1022.2  
##      p_95    p_99 skewness kurtosis      iqr      range_98      range_80  
## 1 1031.8 1037.8 0.1264733 3.539679 10.2 [997.5, 1037.8] [1007.5, 1028.2]
```

Centrálna poloha dát

Hodnota výberového mediánu je 1016.9 hPa. Hodnota výberového priemeru je 1017.3 hPa. Hodnoty sú takmer rovnaké. Priemer je teda dobrý ukazovateľ a dátá zjavne nie sú ovplyvnené veľkým množstvom outlierov ale ukazovatele centrálnej polohy sú veľmi silné.

Modus - najčastejšia hodnota je 1017 hPa.

```
getmode(na.omit(all_data$SLP)) %>%  
  print(cat("Modus: " ))  
  
## Modus: [1] 1017  
median(all_data$SLP, na.rm = TRUE) %>%  
  print(cat("Median: "))
```

```

## Median: [1] 1016.9
mean(all_data$SLP, na.rm = TRUE)%>%
  print(cat("Mean: "))

## Mean: [1] 1017.295
var(all_data$SLP, na.rm = T) %>% print(cat("Rozptyl: ")) # rozptyl

## Rozptyl: [1] 67.58904
var_rozpatie <- max(all_data$SLP, na.rm = T) - min(all_data$SLP, na.rm = T) # variacne rozpatie
print(cat("Variačné rozpätie", var_rozpatie))

## Variačné rozpätie 78.8NULL
# Interquartile range and outliers
Q1 <- quantile(all_data$SLP, 0.25, na.rm = T) # 25% hodnot je mensich a 75% vacsich
Q3 <- quantile(all_data$SLP, 0.75, na.rm = T) # 75% hodnot je mensich a 25% vacsich
IQR <- IQR(all_data$SLP, na.rm = T) # interquartile range

(IQR/2) %>% # interquartile range
  print(cat("Medzikvartilová odchýlka: "))

## Medzikvartilová odchýlka: [1] 5.1
# odlahle hodnoty
length(which(all_data$SLP < (Q1 - 1.5*IQR)))

## [1] 1526
length(which(all_data$SLP > (Q3 + 1.5*IQR)))

## [1] 2066
# extremne hodnoty
length(which(all_data$SLP < (Q1 - 3*IQR)))

## [1] 45
length(which(all_data$SLP > (Q3 + 3*IQR)))

## [1] 0

```

Variabilita

Výberový rozptyl je 67.58904.

Výberová smerodajná odchýlka je 8.221255. To znamená, že hodnoty atmosférického tlaku sa pohybujú približne v rozsahu 8.221255 okolo priemeru.

Variačný koeficient je 0.008081483 alebo v percentách 0,8%, teda hodnoty atmosférického tlaku nie sú variabilné ale naopak sú relatívne nakope.

Variačné rozpätie je 78.8. Daná hodnota predstavuje rozdiel medzi maximálnou a minimálnou nameranou hodnotou atmosférického tlaku. Maximálna hodnota atm. tlaku 1049.4 a minimálna 970.6 hPa. Vychádza to pravdaže z podstaty atmosferického tlaku a jednotiek, v ktorých bol nameraný.

Medzikvartilová odchýlka (IQR/2) je 5.1, teda hodnoty sú rozptýlené približne 5.1 okolo mediánu. Je to o dosť menšie číslo a hovorí nám to o tom, že veľká väčšina dát sa nachádza nakope okolo strednej hodnoty.

Hodnota prvého a tretieho kvartílu je 1012, resp. 1022. Polovica dát sa teda nachádza v rozpäti iba 10 hPa.

Asimetria

Hodnota šikmosti je kladná 0.1264733, ale blízka 0, teda dátu nie sú veľmi zošikmené a ide o prevažne symetrické rozdelenie okolo strednej hodnoty, ktoré je len mierne zošikmené doľava.

Hodnota špicatosti je 3.539679, je o niečo väčšia ako 3 teda hodnoty majú špicatejšie rozdelenie. To znamená, že v súbore sa nachádza viac hodnôt bližších k strednej hodnote.

Boxplot

Hodnota 3.kvartilu je 1022.2 a hodnota 1. kvartilu 1012, teda medzikvartilové rozpätie (IQR) je 10.2. Uprostred krabice je zvýraznený medián hrubou čierou (1017). Medián sa nachádza v strede krabice, teda vyzerá, že dátu sú symetrické okolo strednej hodnoty.

Ďalej z boxplotu vidieť maximálnu a minimálnu hodnotu (vonkajšie hradby boxplotu). Maximálna hodnota (1037.5) je vypočítaná ako 3.kvartil + 1.5 * IQR (medzikvartilové rozpätie). Minimálna hodnota (996.7) je vypočítaná ako 1.kvartil - 1.5 * IQR (medzikvartilové rozpätie).

Všetky hodnoty nachádzajúce sa nad a pod maximálnou a minimálnou hodnotou môžme považovať za odľahlé hodnoty. Počet odľahlých hodnôt nad maximálnou hodnotou je 2066, pod minimálnou sa nachádza 740 hodnôt.

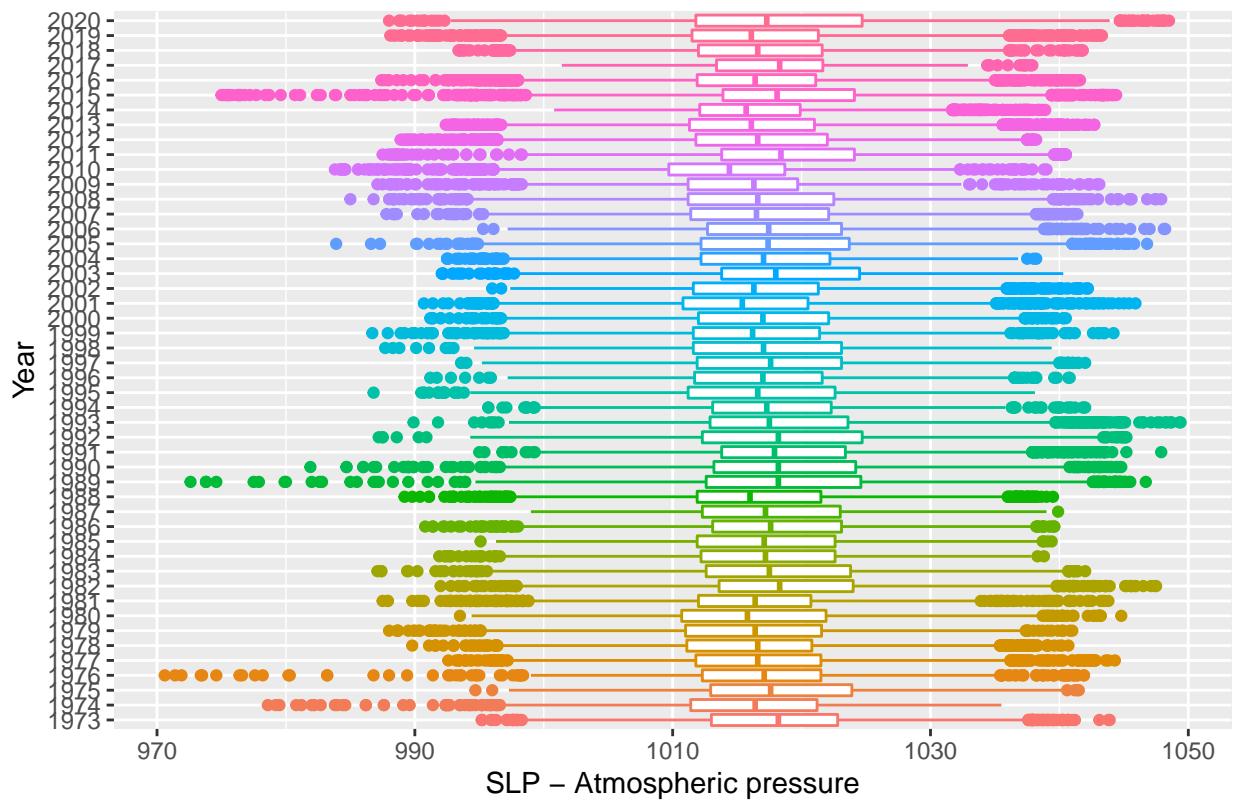
Nakoniec pre odľahlé hodnoty overíme či patria medzi extrémne. Horná hranica extrémnych hodnôt je vypočítaná ako 3.kvartil + 3 * IQR. Dolná hranica extrémnych hodnôt je vypočítaná ako 1.kvartil - 3 * IQR. V dátach sa nachádza 45 extrémne nízkych hodnôt, ktoré sú nižšie ako hodnota 1.kvartilu - 3 * IQR, teda nižšie ako 981.4 hPa. Extrémne vysoké hodnoty sa v dátach nenachádzajú.

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

ggplot(data = df, aes( SLP,factor(year), colour=year)) +
  geom_boxplot() +
  labs(title = paste("Boxplot atmosférického tlaku, jednotlivé roky")) +
  xlab("SLP - Atmospheric pressure") +
  ylab("Year") +
  theme(legend.position = "none")

## Warning: Removed 224734 rows containing non-finite values (stat_boxplot).
```

Boxplot atmosférického tlaku, jednotlivé roky

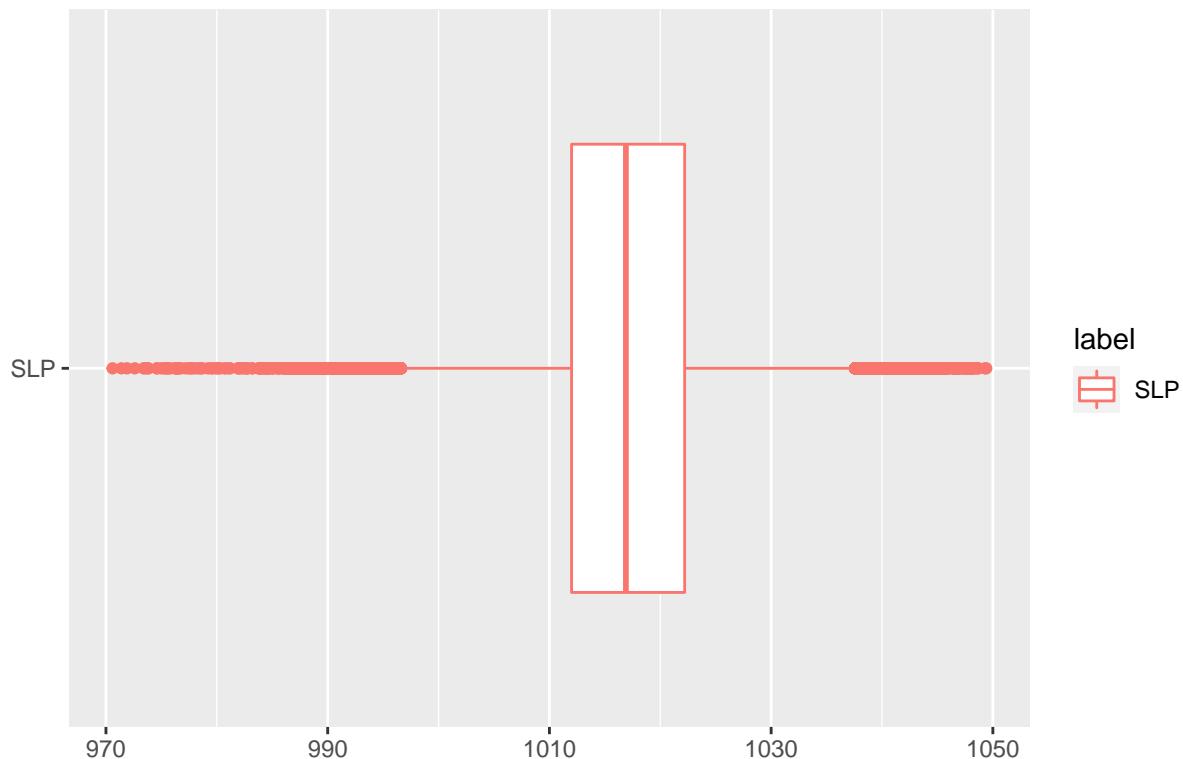


```
df <- all_data %>%
  dplyr::select('SLP') %>%
  tidyr::gather(key='label', value = 'pressure')

ggplot(data = df, aes( pressure,factor(label), colour=label)) +
  geom_boxplot() +
  labs(title = paste("Boxplot atmosferickeho tlaku")) +
  xlab("") +
  ylab("")
```

Warning: Removed 224734 rows containing non-finite values (stat_boxplot).

Boxplot atmosferickeho tlaku



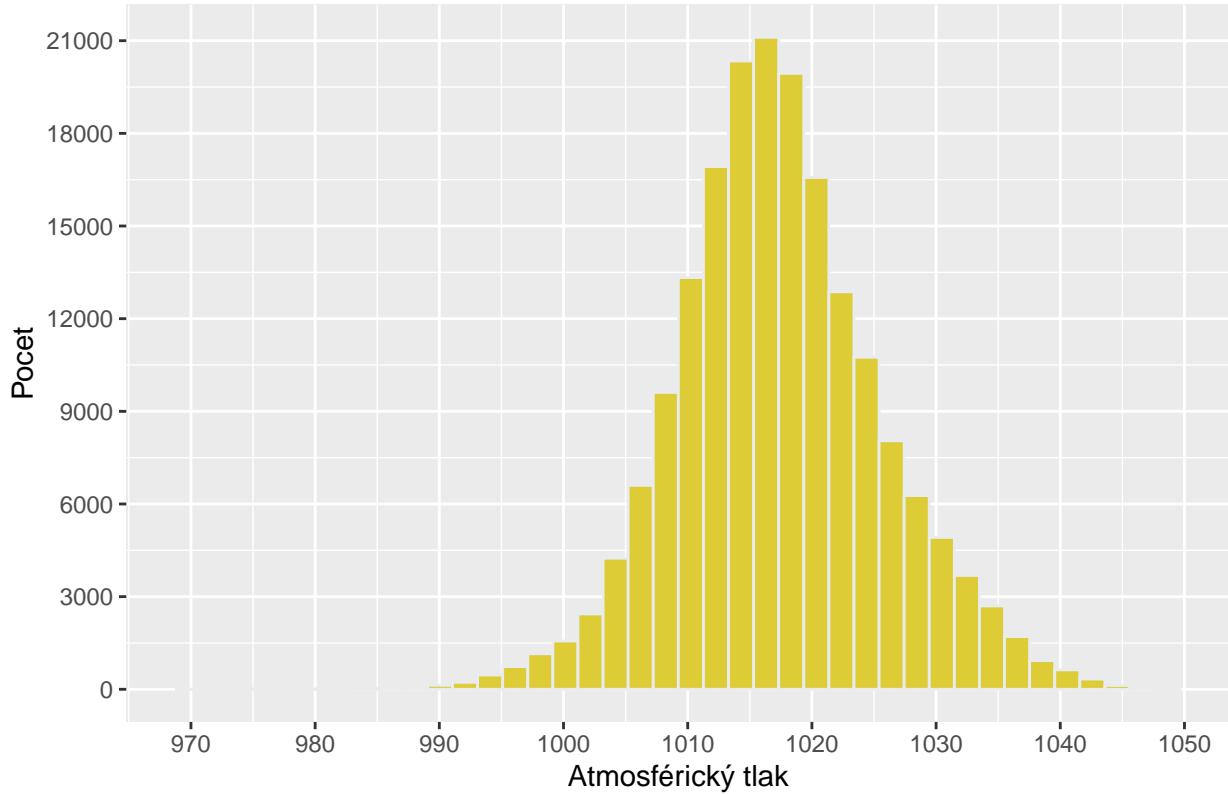
Histogram

Aj z histogramu vidíme, že rozdelenie hodnôt sa podobá na normálne, pričom najpočetnejšie sú hodnoty okolo 1010 - 1020 hPa.

```
ggplot(all_data, aes(x = SLP)) +  
  geom_histogram(bins = 40, fill = "#ddcc36", color = "#e9ecef") +  
  labs(title = paste("Sea level pressure histogram")) +  
  xlab("Atmosférický tlak") +  
  ylab("Počet") +  
  scale_x_continuous(breaks = seq(950, 1200, by = 10)) +  
  scale_y_continuous(breaks = seq(0, 60000, by = 3000))
```

```
## Warning: Removed 224734 rows containing non-finite values (stat_bin).
```

Sea level pressure histogram



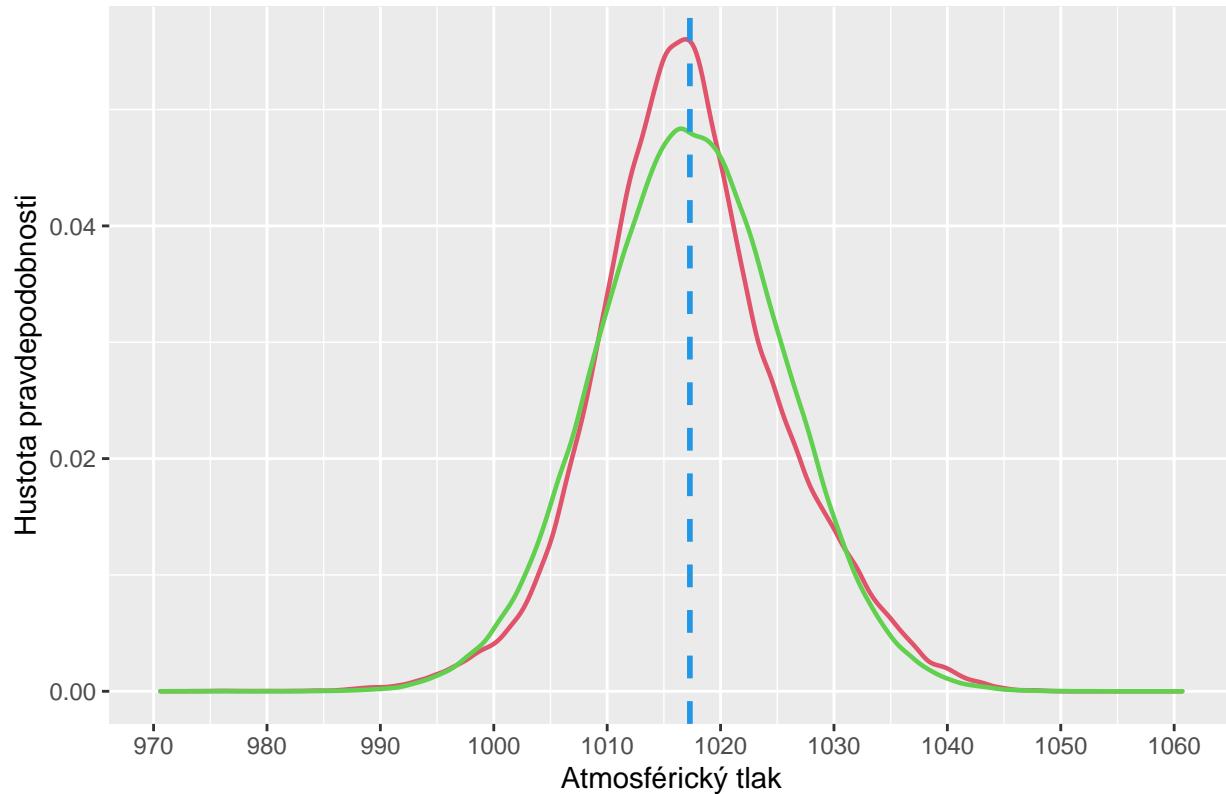
Graf hustoty

Graf hustoty slúži na porovnanie priebehu hustoty pravdepodobnosti normálneho rozdelenia (zelená čiara) a odhadu hustoty vypočítaného z namerných hodnôt atmosférického tlaku (červená čiara). Čiary sú približne rovnaké, teda ide o normálne rozdelenie, ktoré je ale trocha špicatejšie. Modrá prerušovaná čiara predstavuje priemernú hodnotu atmosférického tlaku.

```
# denisty plot
# data z normalneho rozdelenia
data_norm <- data.frame(dens = c(rnorm(length(na.omit(all_data$SLP)), mean(all_data$SLP, na.rm = T), sd = 1), 0))

# porovnanie hodnot normalneho rozdelenia a SLP
ggplot(all_data, aes(x = SLP), 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(SLP, na.rm = T)),
             color = 4, linetype = "dashed", size = 1) +
  scale_x_continuous(breaks = seq(900, 1100, by = 10)) +
  labs(title = paste("Odhad hustoty hodnôt atmosférického tlaku")) +
  xlab("Atmosférický tlak") +
  ylab("Hustota pravdepodobnosti")
```

Odhad hustoty hodnôt atmosférického tlaku



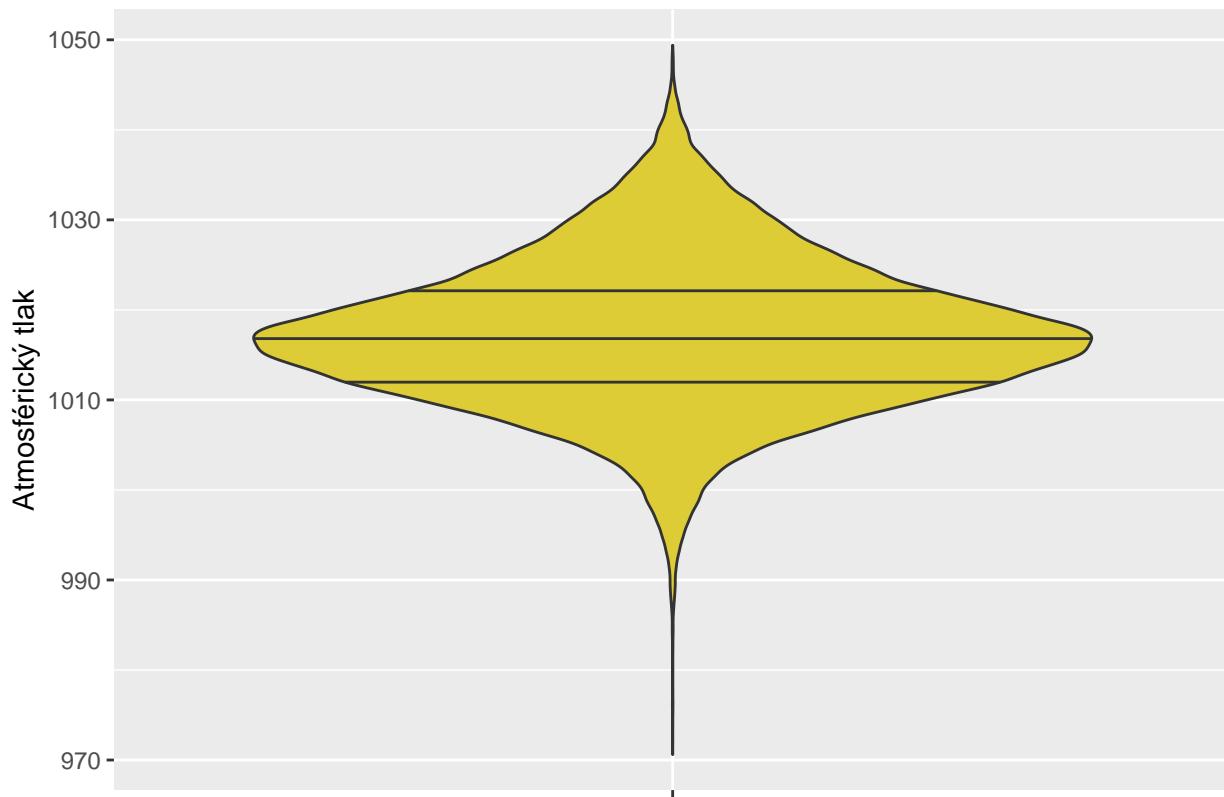
Husľový graf

Husľový graf doplnený o hlavné kvartily zobrazuje rozdelenie hustoty, pričom aj podľa tohto grafu vidíme, že ide o normálne rozdelenie. Dáta sú najpočetnejšie v okolí hodnoty 1017, postupne smerom k vyšším a nižším hodnotám sa ich hustota zmenšuje.

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

ggplot(data = df, aes(factor(label), slp, fill = slp)) +
  geom_violin(draw_quantiles = c(0.25, 0.5, 0.75), fill = "#ddcc36") +
  labs(title = paste("Husľový graf atmosférického tlaku"), y = "Atmosférický tlak", fill = "Sea level p...") +
  theme(axis.title.x = element_blank()) +
  theme(axis.text.x = element_blank())
```

Huslový graf atmosférického tlaku



Q-Q graf

Graf zobrazuje odchýlku empirického od teoretického normálneho rozdelenia. Empirické rozdelenie je v našom prípade rozdelenie nameraných hodnôt atmosférického tlaku. Kedže body ležia blízko priamky normálneho rozdelenia, môžeme povedať, že rozdelenie hodnôt je normálne. Mierne sa odchyľuje len v oblasti dolných kvantilov.

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

Q–Q graf atmosférického tlaku

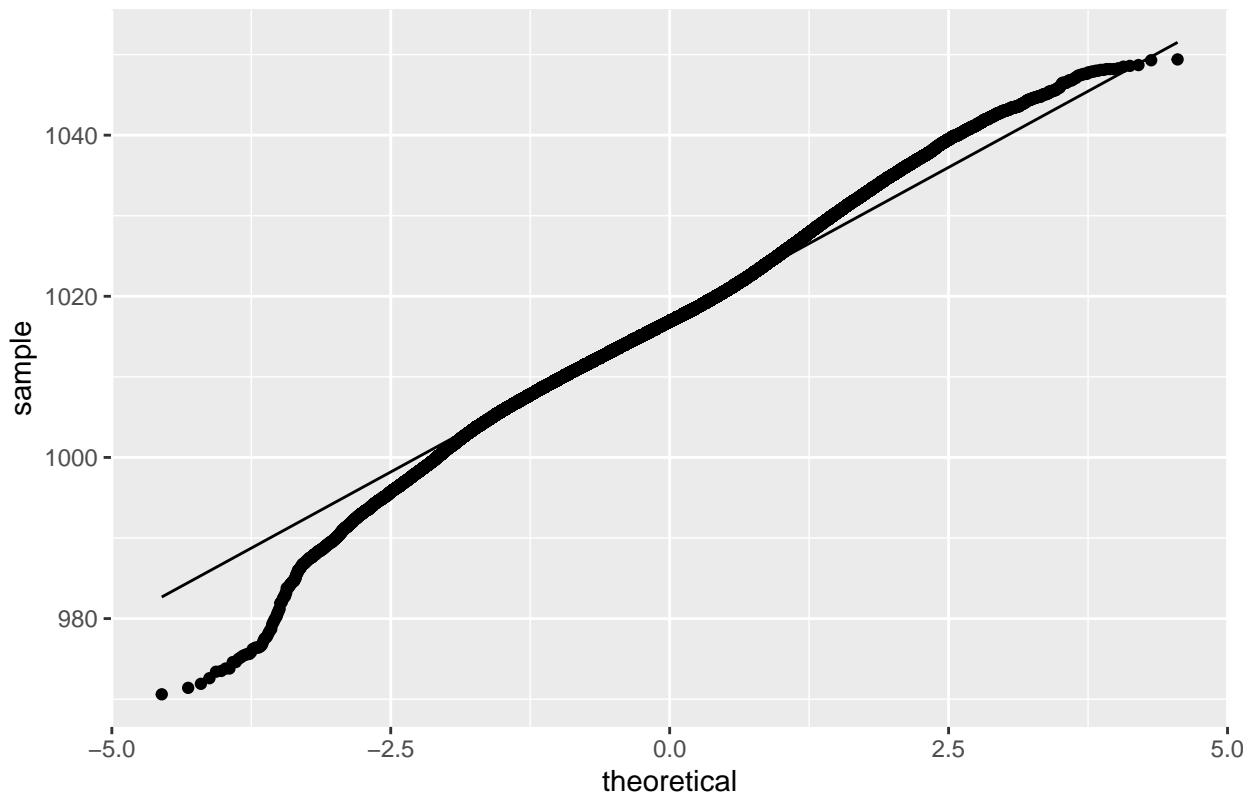


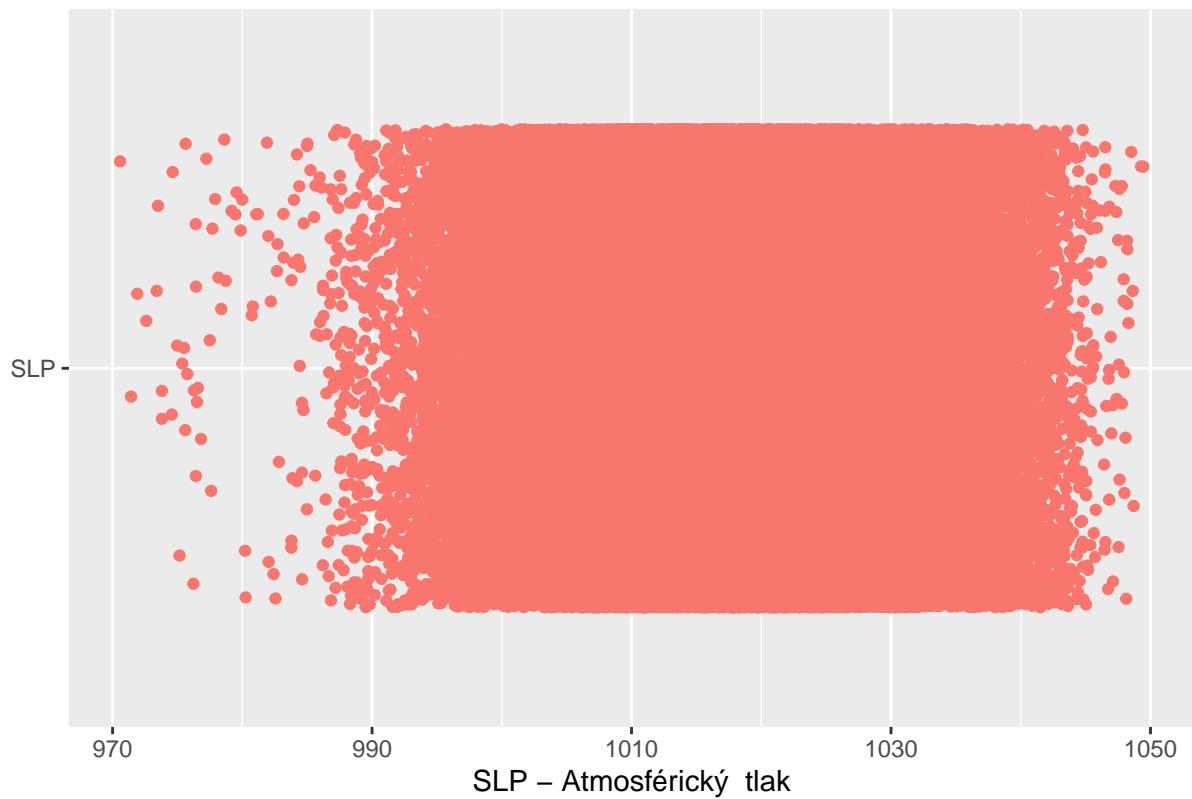
Diagram rozptylenia

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

ggplot(data = df, aes( pressure,factor(label), colour=label)) +
  geom_jitter() +
  labs(title = paste("Diagram rozptylenia pre atmosférický tlak")) +
  xlab("SLP – Atmosférický tlak") +
  ylab("") +
  theme(legend.position = "none")

## Warning: Removed 224734 rows containing missing values (geom_point).
```

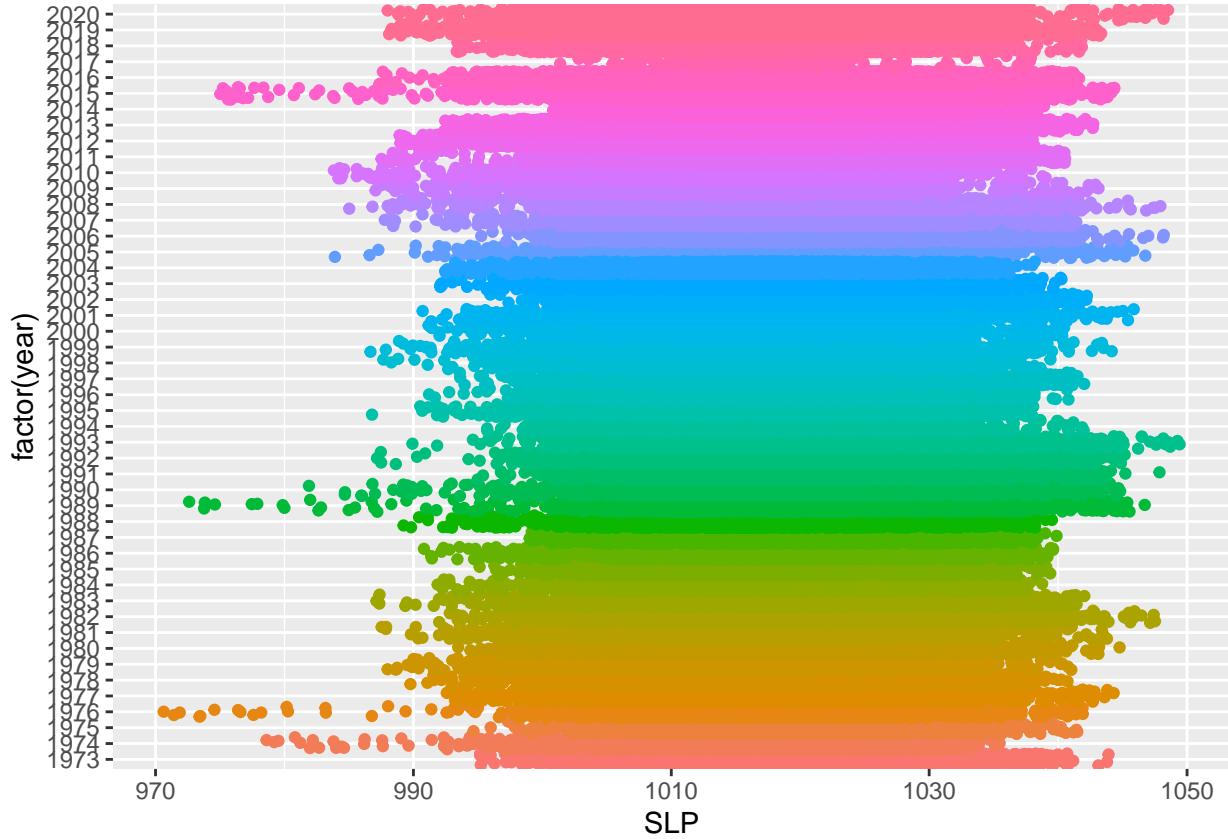
Diagram rozptylenia pre atmosférický tlak



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

ggplot(data = df, aes( SLP,factor(year), colour=year)) +
  geom_jitter() +
  theme(legend.position = "none")

## Warning: Removed 224734 rows containing missing values (geom_point).
```



Graf polosum

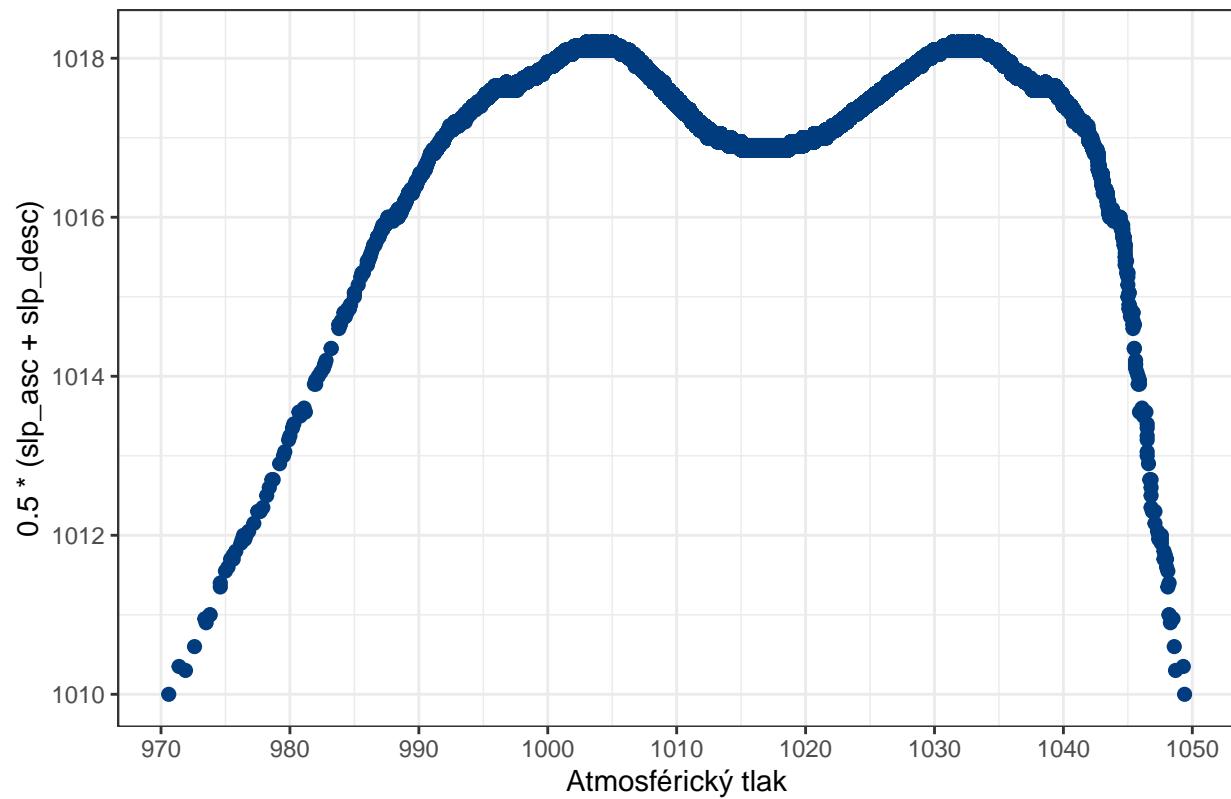
Z grafu vidno, že hodnoty sú takmer symetrické okolo mediánu.

```

slp <- all_data$SLP
slp_asc <- sort(slp, decreasing = FALSE)
slp_desc <- sort(slp, decreasing = TRUE)

ggplot(data.frame(slp_asc), aes(x = slp_asc, y = 0.5*(slp_asc + slp_desc))) +
  geom_point(size = 2, color = "#013c7f") +
  scale_x_continuous(breaks = seq(0, 2000, by = 10)) +
  labs(title = "Graf polosum pre atmosférický tlak", x = "Atmosférický tlak") +
  theme_bw()
  
```

Graf polosum pre atmosférický tlak

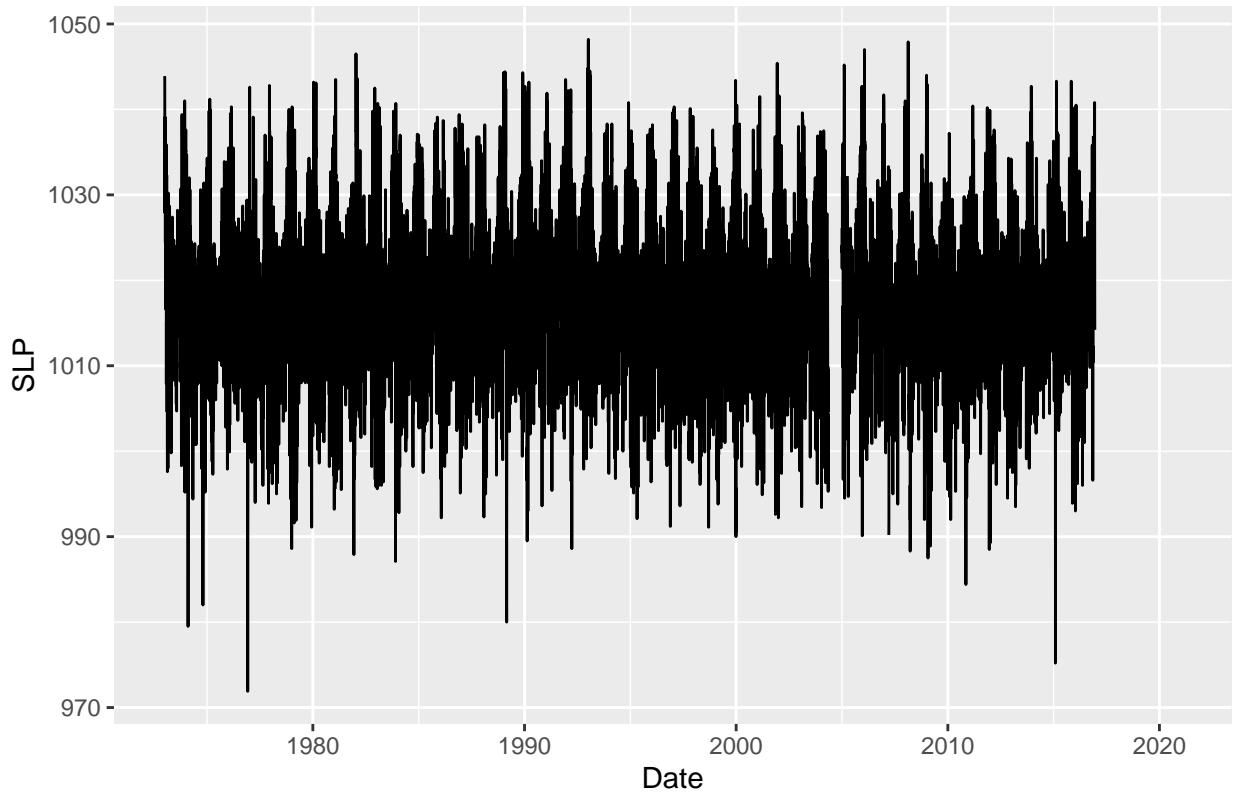


Časový graf atmosférického tlaku

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

## Warning: Removed 1480 row(s) containing missing values (geom_path).
```

Time graph of sea level pressure



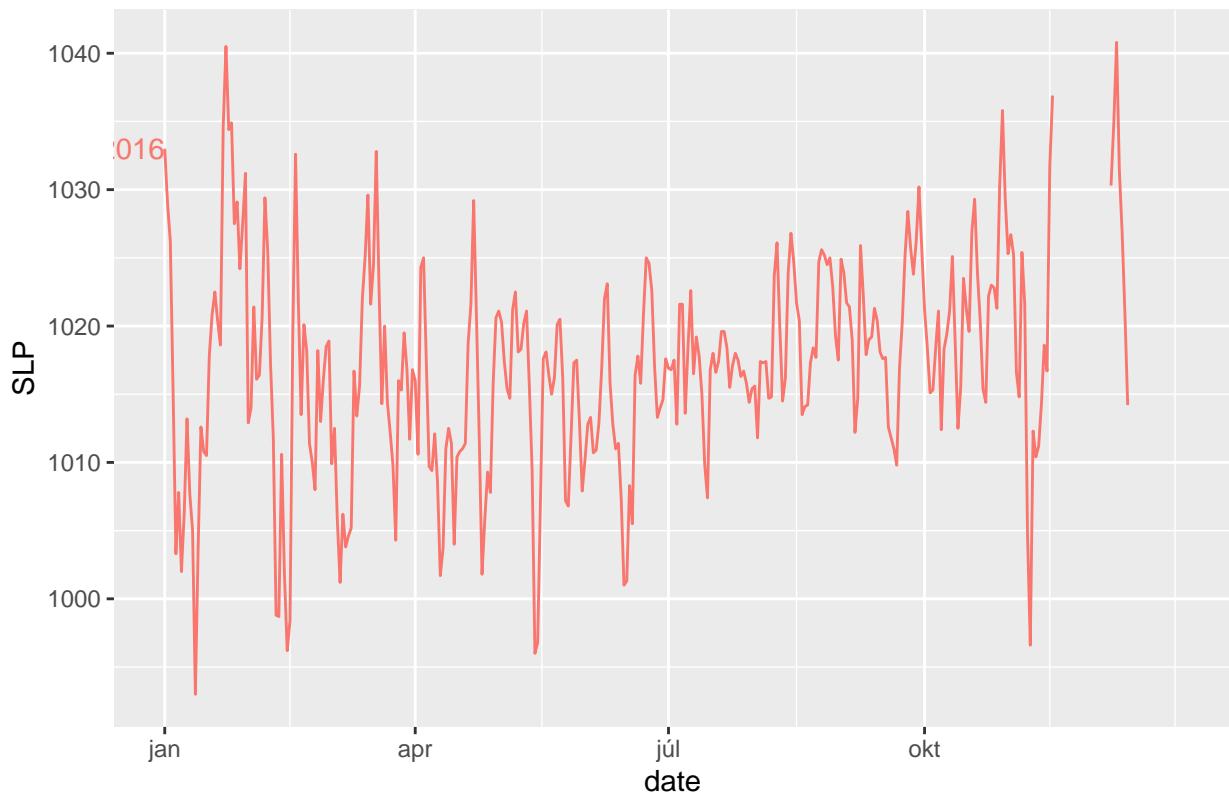
```

all_data %>%
  dplyr::mutate(
    date = as_date(DATE)
  ) %>%
  dplyr::distinct(date, .keep_all=TRUE) %>%
  dplyr::select(date, SLP) %>%
  as_tsibble(
    index = date
  ) %>%
  tsibble::fill_gaps() %>%
  dplyr::filter(year(date)>2015) %>%
  gg_season(SLP, labels = "both") +
  labs(y = "SLP",
       title = "Seasonal plot: SLP")

## Warning: Removed 1480 row(s) containing missing values (geom_path).
## Warning: Removed 9 rows containing missing values (geom_text).

```

Seasonal plot: SLP



```
all_data %>%
  dplyr::mutate(
    year_month = yearmonth(DATE)
  ) %>%
  dplyr::group_by(year_month) %>%
  dplyr::summarise(SLP = na.omit(mean(SLP))) %>%
  as_tsibble(
    index = year_month
  ) %>%
  tsibble::fill_gaps() %>%
  dplyr::filter(year(year_month)>210) %>%
  gg_subseries(SLP, period = "1 year") +
  labs(y = "SLP",
       title = "Seasonal plot: SLP")

## `summarise()` has grouped output by 'year_month'. You can override using the `groups` argument.
## Warning: Removed 3 row(s) containing missing values (geom_path).
```

Seasonal plot: SLP

