LP

Adam Štuller

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

Liquid precipitation

Spracovanie stĺpca

Tieto data treba najprv spracovať. Zrážky sa nachádzali pôvodne v dvoch stĺpcoch. Jeden obsahoval merania raz za 24 hodin. Druhy obsahoval spolu merania za 12 a 6 hodin. Nemohli sme ale použiť iba jedno z týchto meraní, pretože v každnom chýbali nejaké časové úseky. Preto sme sa rozhodli pospájať jednotlivé merania dokopy.

Pracovali sme s množstvom zrážok na jeden deň. 24 hodinové merania sú na celý đen, s nimi sme nerobili nič. 6 a 12 hodinové merania sme najprv zosumovali pre každý deň. Potom sme spojili jednotlivé dni pomocou funkcie coalescence, tak že, ak bolo k dispozícií 12 hodinové meranie tak sme zobrali to. Ak nebolo tak sme skúsili 6 hodinové meranie a ak nebolo ani to tak sme vzali 24 hodinové meranie. Defaultna hodnota bola 0.

```
all_data %>%
  dplyr::mutate(
    date = as_date(DATE)
  ) %>%
  select(date, LP, LP24) %>%
  separate(LP, c('lp_observation_period', 'lp_observation', NA, NA)) %>%
  filter(lp_observation_period == 12) %>%
  dplyr::mutate(lp_observation = map_dbl(lp_observation, process_col, 10)) %>%
  dplyr::select(date, lp_observation) %>%
  dplyr::group_by(date) %>%
  dplyr::summarise(LP12 = sum(lp_observation)) %>%
  as_tsibble(
    index = date
  ) %>%
  dplyr::filter(year(date)>0) %>%
  tsibble::fill_gaps() -> df_lp12


all_data %>%
  dplyr::mutate(
    date = as_date(DATE)
  ) %>%
  select(date, LP, LP24) %>%
  separate(LP, c('lp_observation_period', 'lp_observation', NA, NA)) %>%
  filter(lp_observation_period == "06") %>%
  dplyr::mutate(lp_observation = map_dbl(lp_observation, process_col, 10)) %>%
  dplyr::select(date, lp_observation) %>%
  dplyr::group_by(date) %>%
```

```
  dplyr::summarise(LP6 = sum(lp_observation)) %>%
  as_tsibble(
    index = date
  ) %>%
  dplyr::filter(year(date)>0) %>%
  tsibble::fill_gaps() -> df_lp6

all_data %>%
  dplyr::mutate(
    date = as_date(DATE)
  ) %>%
  select(date, LP24) %>%
  distinct(date, .keep_all = TRUE) %>%
  as_tsibble(
    index = date
  ) %>%
  tsibble::fill_gaps() -> df_lp24

merge(df_lp6, df_lp12, by = "date", all = TRUE) %>%
  merge(df_lp24, by = "date", all = TRUE) -> merged_df

merged_df %>%
  dplyr::mutate(
    LP = coalesce(LP12, LP6, LP24) %>% replace_na(0)
  ) %>%
  as_tsibble(
    index = date
  ) -> lp_df
```

```
describe(lp_df$LP)
```

```
## lp_df$LP
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    17532        0      254    0.664    1.433    2.546      0.0      0.0
##      .25      .50      .75      .90      .95
##      0.0      0.0      0.5      4.1      8.6
##
## lowest :   0.0   0.1   0.2   0.3   0.3, highest:  66.0  68.1  82.0  90.0 248.3
```

Centralna poloha dát

Hodnota vyberoveho medianu je 0, modus je 0 a vyberovy priemer je 1.432945. Znamená to, že vačšina dát je bud priamo 0, teda v tom dni nepršalo alebo je veľmi blížka nule, teda pršalo iba mierne.

```
getmode(na.omit(lp_df$LP)) %>%
  print(cat("Modus: " ))
```

```
## Modus: [1] 0
```

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

```
## Median: [1] 0
```

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

```
## Mean: [1] 1.432945
```

Variabilita

Vyberovy rozptyl je 25.39282 . Variancny koeficient je 3.516623. Dáta sú teda relatívne s veľkou variabilitou.

Variacne rozpatie je 248.3 teda rozdiel medzi najmensim a najvacsim prvkom je dosť velky. Treba ale podotknuť, že takýchto veľkých hodnôt je tam niekoľko a môže ísť o tzv. storočnú vodu. Väčšina dát sa drží na nižších hodnotách.

Medzikvantilova odchýlka je **é.ľť.** Je to o dost malé cislo a hovori nam to o tom, ze velka vacsina dat sa nachadza nakope okolo stednej hodnoty.

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

```
## Variacne rozpatie 248.3NULL
```

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

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

```
## Medzikvantilova odchýlka: [1] 0.25
```

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

```
## Rozptyl: [1] 25.39282
```

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

```
## Variacny koeficient: [1] 3.516623
```

```
summary(lp_df$LP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   1.433   0.500 248.300
```

```
lp_df$LP %>% profiling_num()
```

```
##   variable     mean  std_dev variation_coef p_01 p_05 p_25 p_50 p_75 p_95 p_99
## 1      var 1.432945 5.039129       3.516623    0    0    0    0  0.5  8.6 21.1
##   skewness kurtosis iqr  range_98 range_80
## 1 17.08225 687.5848 0.5 [0, 21.1] [0, 4.1]
```

Asymetria

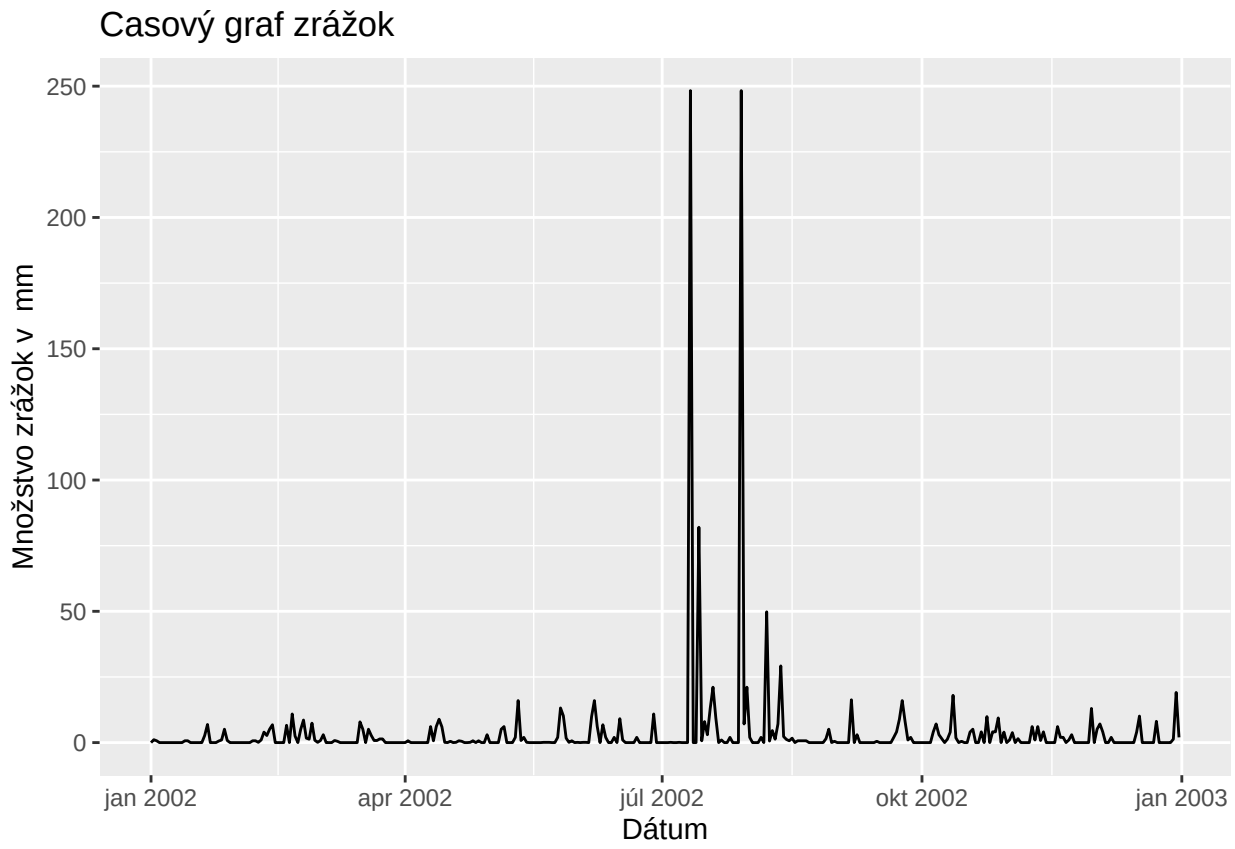
Šikmost (skewness) je 17.08225. Je velmi kladna, teda rozdelenie je poriadne zasikmene do lava.

Špicatost (kurtosis) - 687.5848 je kladna a teda poriadne spicatejsia ako pre data z normálneho rozdelenia.

Časový graf

```
lp_df %>%
  filter(year(date) == 2002 ) %>%
  autoplot( LP) +
```

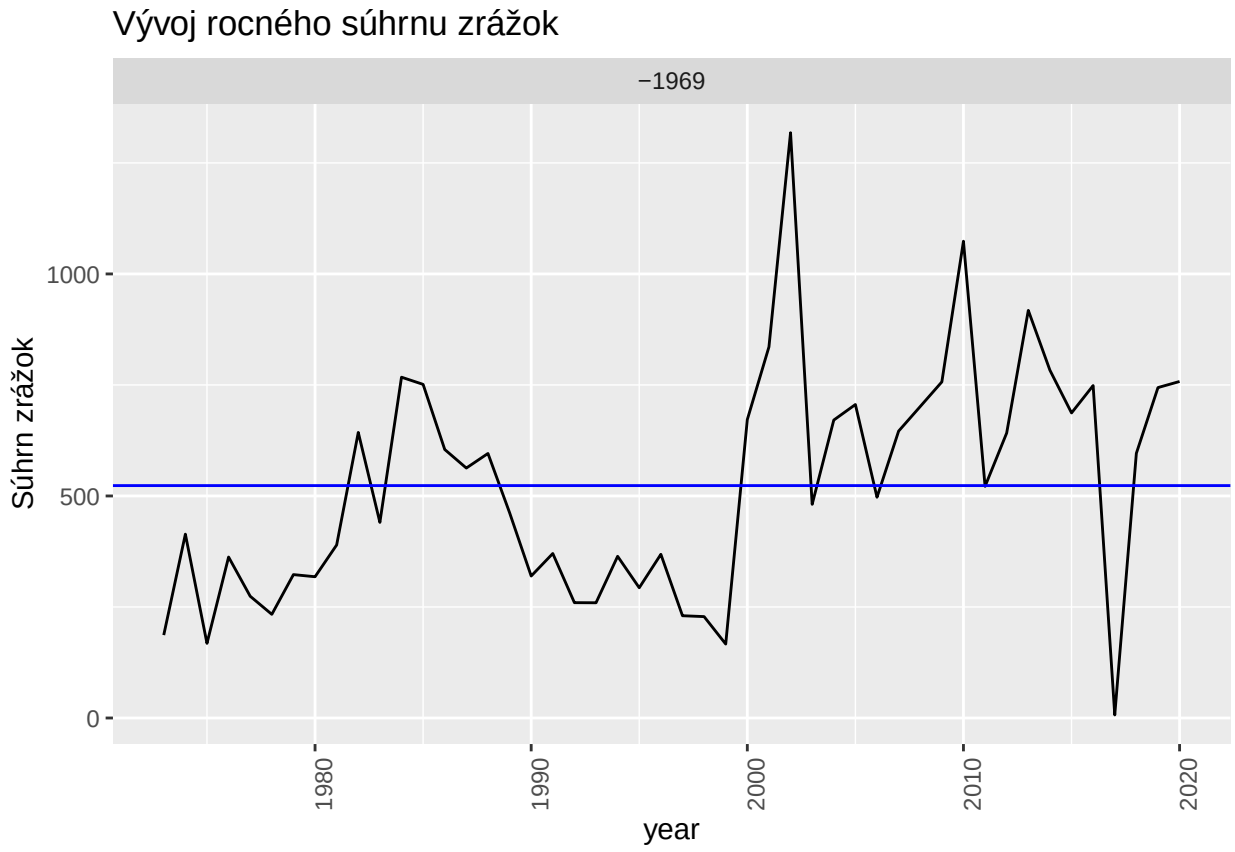
```
    labs(title = "Časový graf zrážok",
         y = "Množstvo zrážok v  mm",
         x = "Dátum"
         )
```



Na časovom grafe vidíme, že iba niekoľko hodnôt je extremnych. Mohli y to kludne byť vychýlené hodnoty ale rovnako aj storočná voda a nám tie dáta veľmi neubližujú a preto ich nebudeme odstraňovať.

Môžeme sa pozrieť aj na ročný súhrn zrážok.

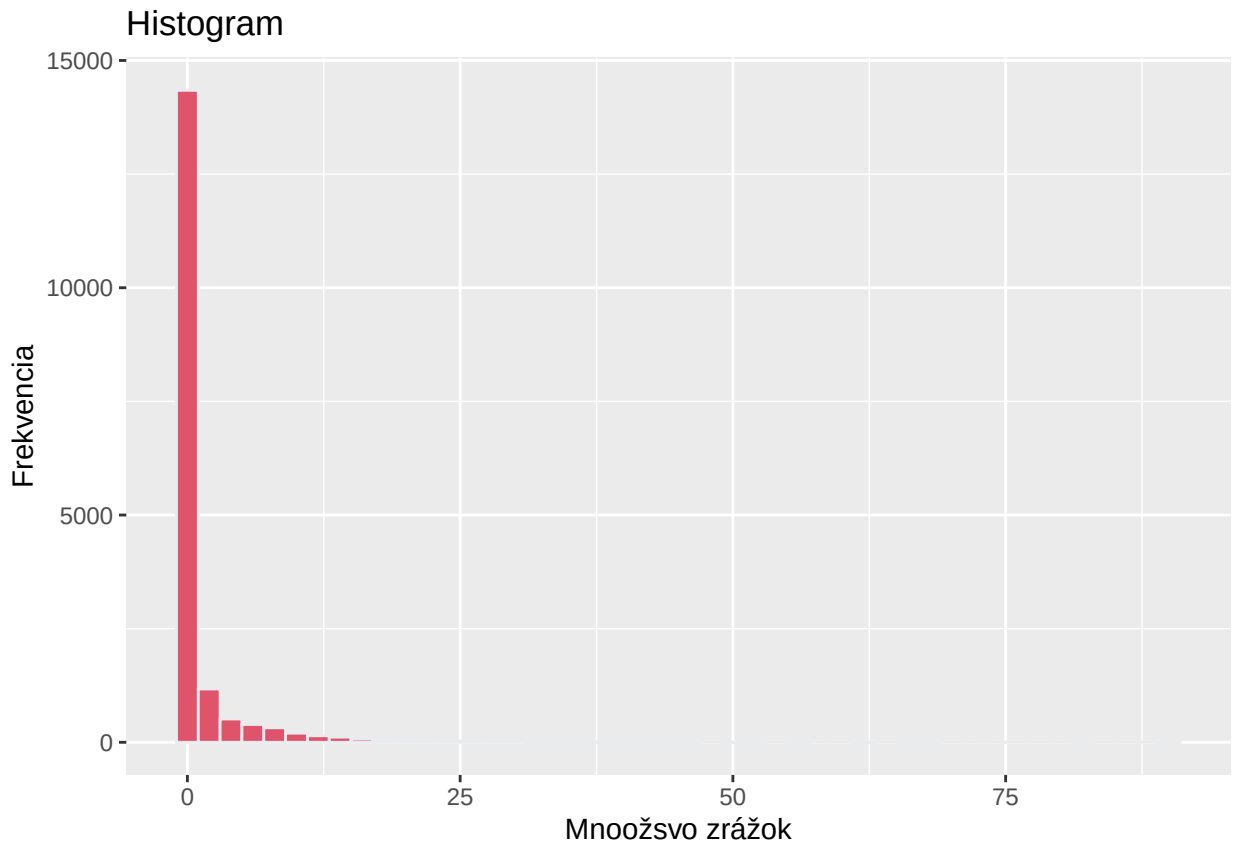
```
lp_df %>%
  as.data.frame() %>%
  dplyr::mutate(
    year = year(date)
  ) %>%
  dplyr::group_by(year) %>%
  dplyr::select(-date) %>%
  dplyr::summarise(LP_SUM = na.omit(sum(LP))) %>%
  as.data.frame() %>%
  distinct(year, .keep_all = TRUE) %>%
  as_tsibble(
    index = year
    ) %>%
  tsibble::fill_gaps() %>%
  gg_subseries(LP_SUM, period = "1 year") +
  labs(y = "Súhrn zrážok",
       title = "Vývoj ročného súhrnu zrážok")
```



Histogram

Na histograme vidíme, že dáta sú z rozdelenia podobného exponencionálnemu. Najviac dát je v okolí nuly. Odstránili sme pre histogram najväčšie hodnoty lebo ho značne roztahovali.

```
lp_df %>%
  filter(LP < 200) %>%
  ggplot( aes(x=LP)) +
    geom_histogram(bins = 40, binwidth = 2,fill="2", color="#e9ecef") +
    labs(title = paste("Histogram")) +
    xlab("Mnoožsvo zrážok") +
    ylab("Frekvencia")
```



Boxplot

Z boxplotu vidíme tú istú informáciu. Veľké množstvo dát sa nachádza okolo nuly a iba niekoľko ma väčšiu hodnotu.

```
df <- lp_df %>%
  dplyr::select('LP') %>%
  tidyr::gather(key='label', value = 'lp')

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

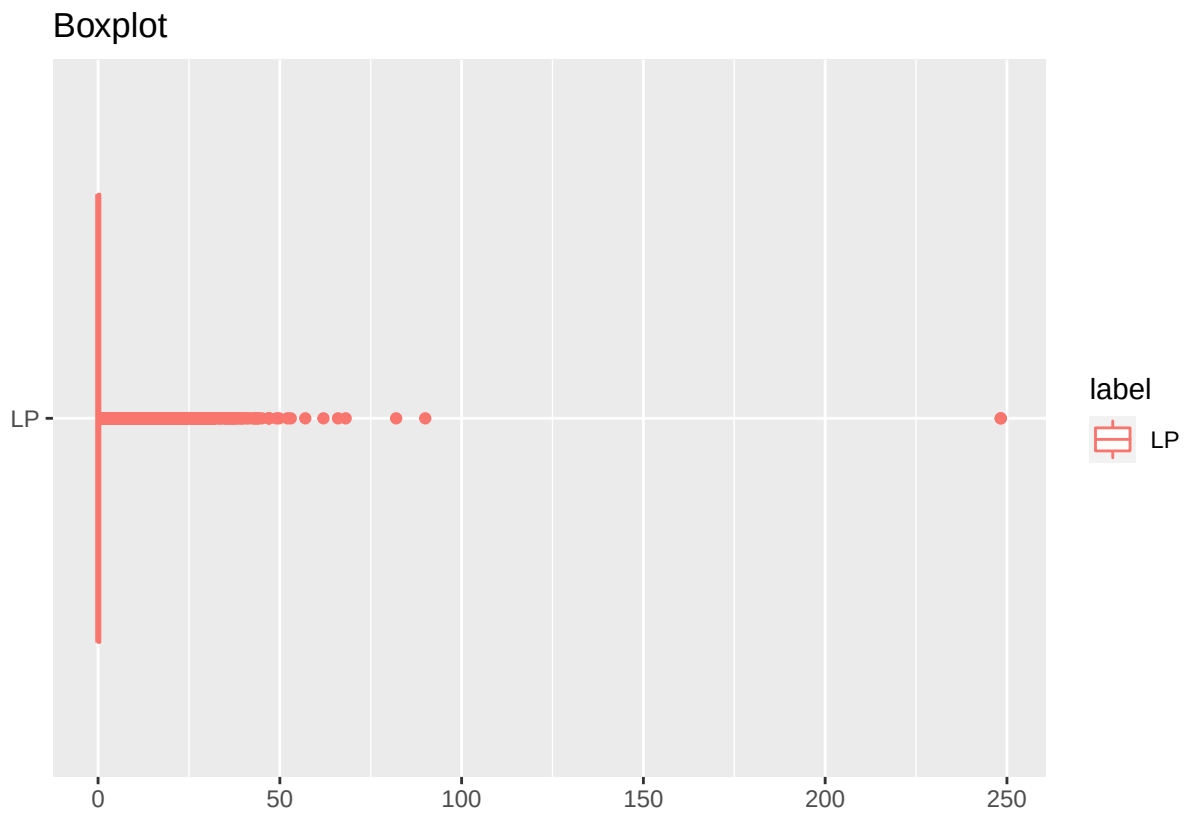
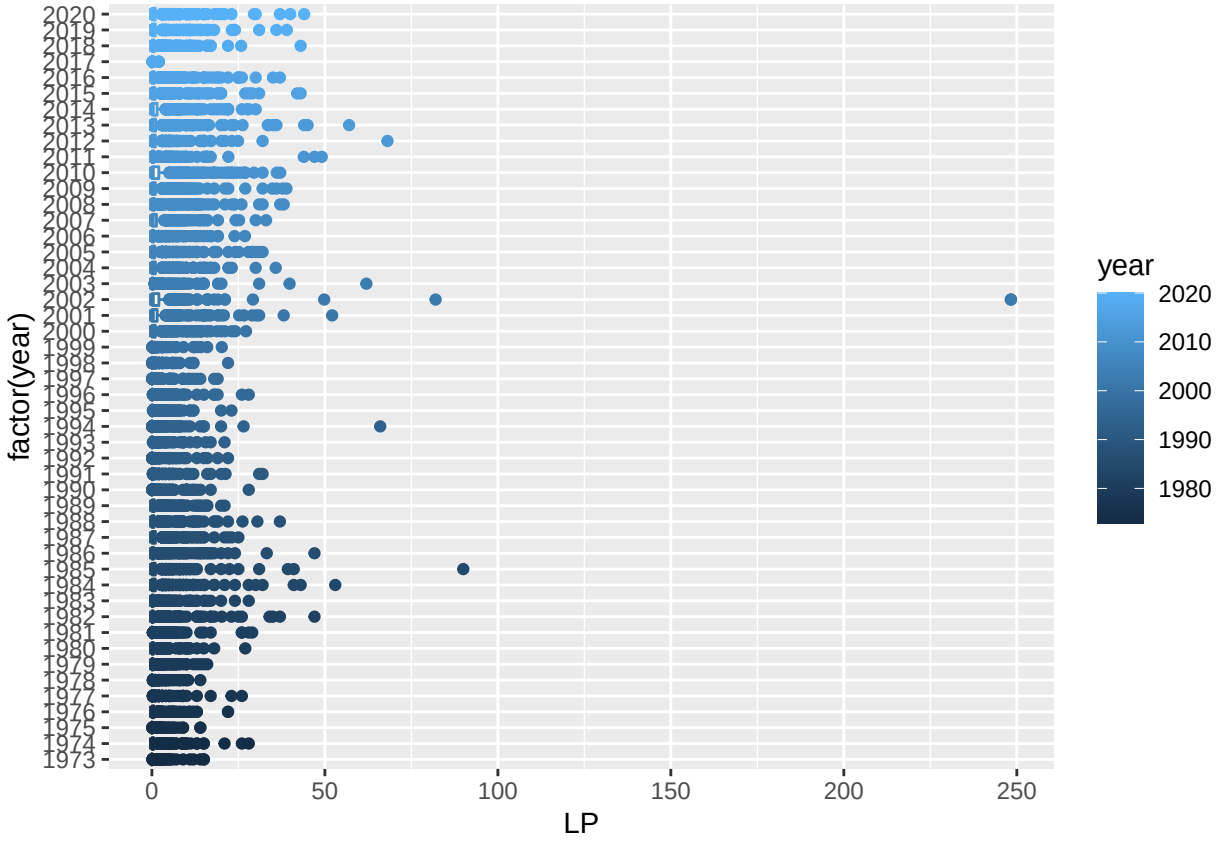


Diagram rozptýlenia pre jednotlivé roky

Vidíme opať, že rok 2017 je čudný.

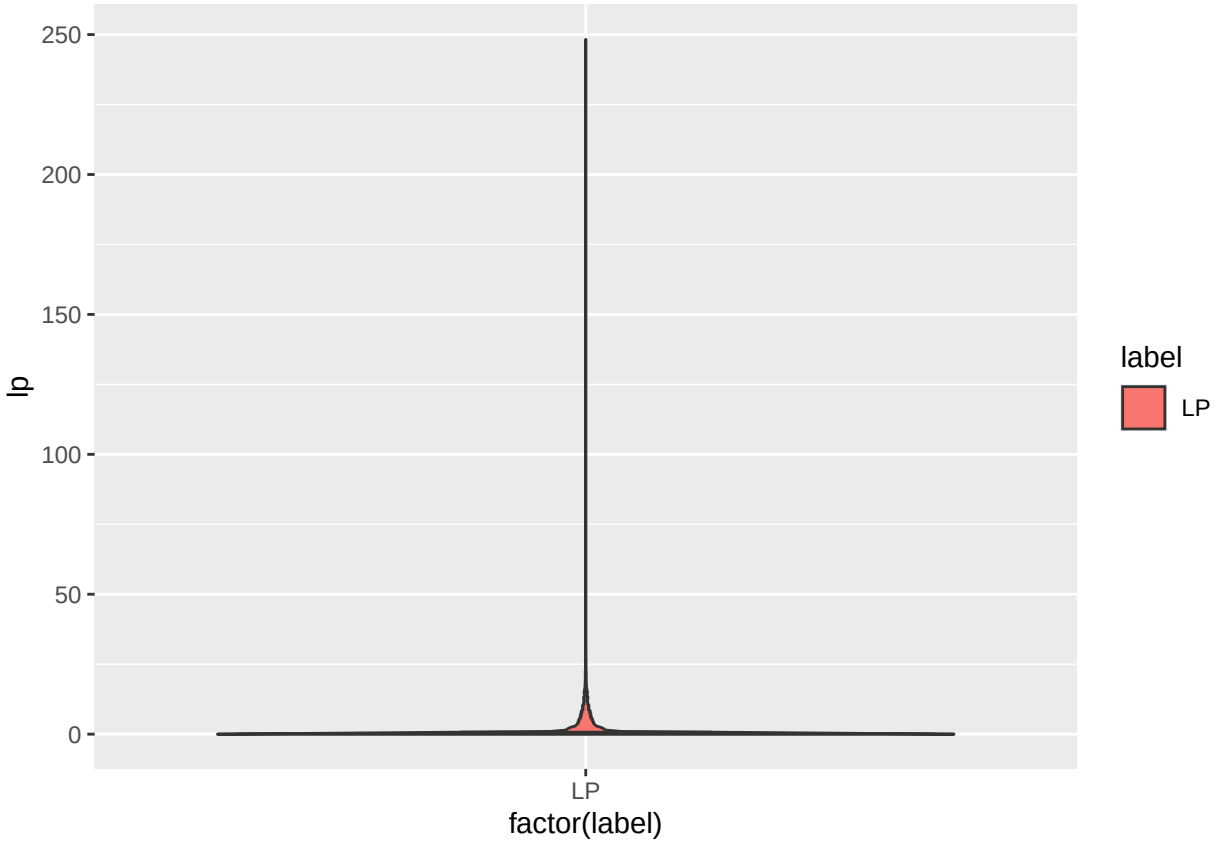
```
df <- lp_df %>%
  dplyr::mutate(
    year = year(date)
  ) %>%
  dplyr::select(all_of(c('year', 'LP')))

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



Violin

```
df <- lp_df %>%
  dplyr::select('LP') %>%
  tidyr::gather(key='label', value = 'lp')

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

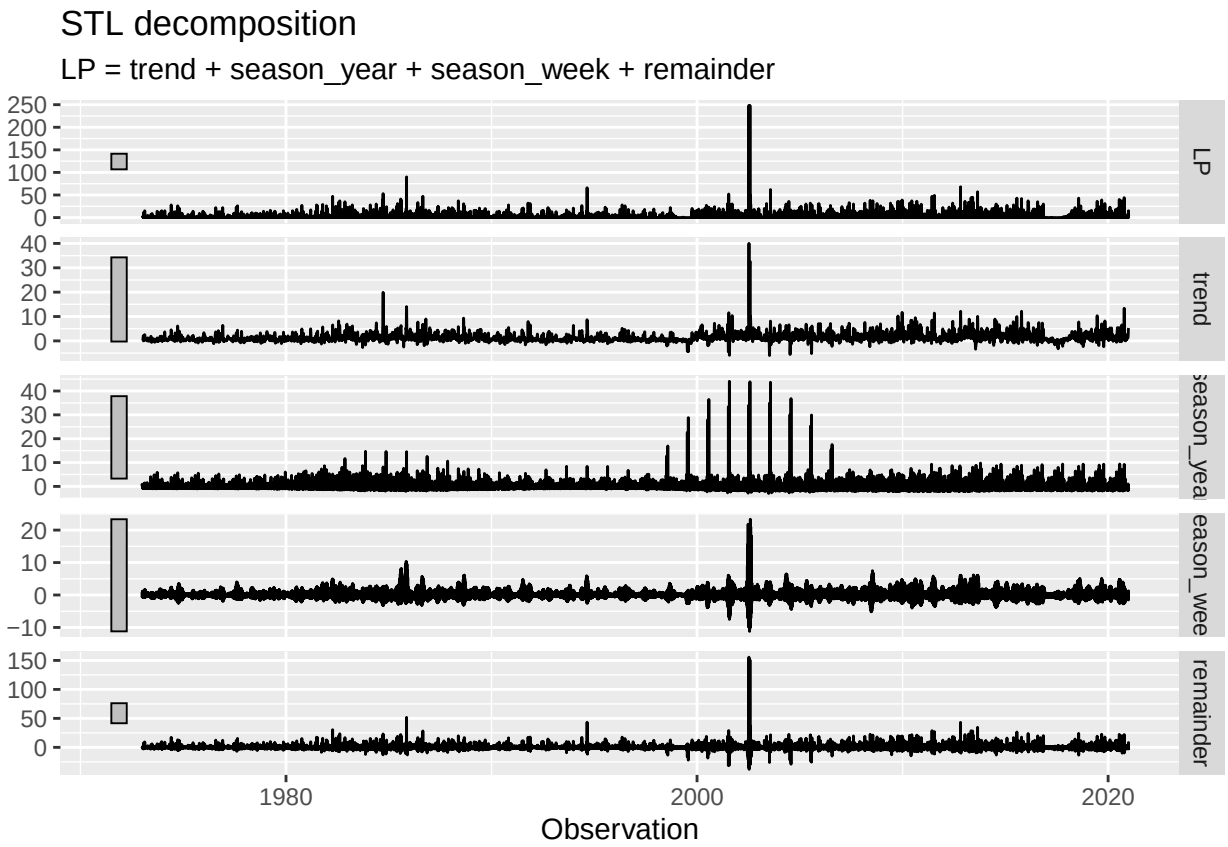
```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```



Dekompozícia

Pri dekompozícií vídíme, že existuje silný sezónny prírastok týchto časových radov.

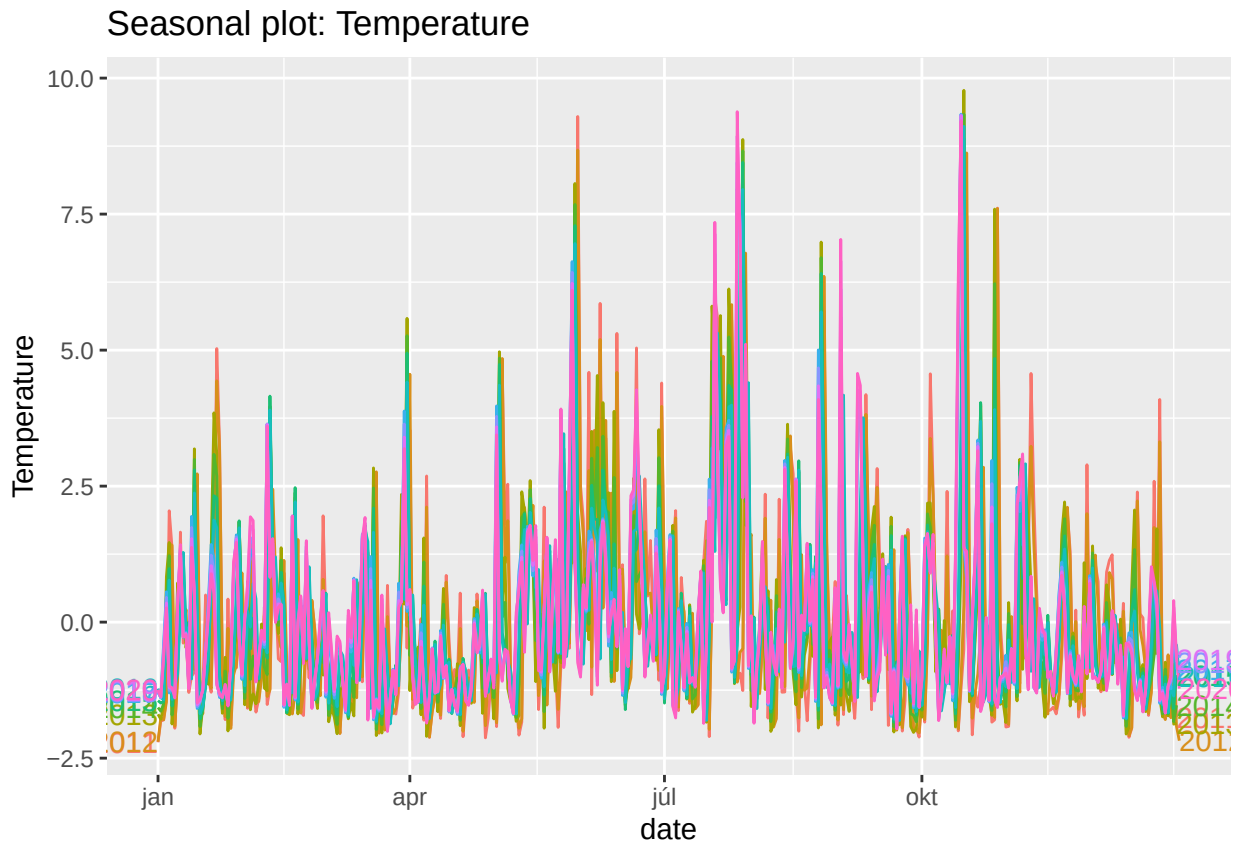
```
lp_df %>%
  model(STL(LP )) -> m

m %>%
  components() %>%
  autoplot() + labs(x = "Observation")
```



Sezónny príspevok

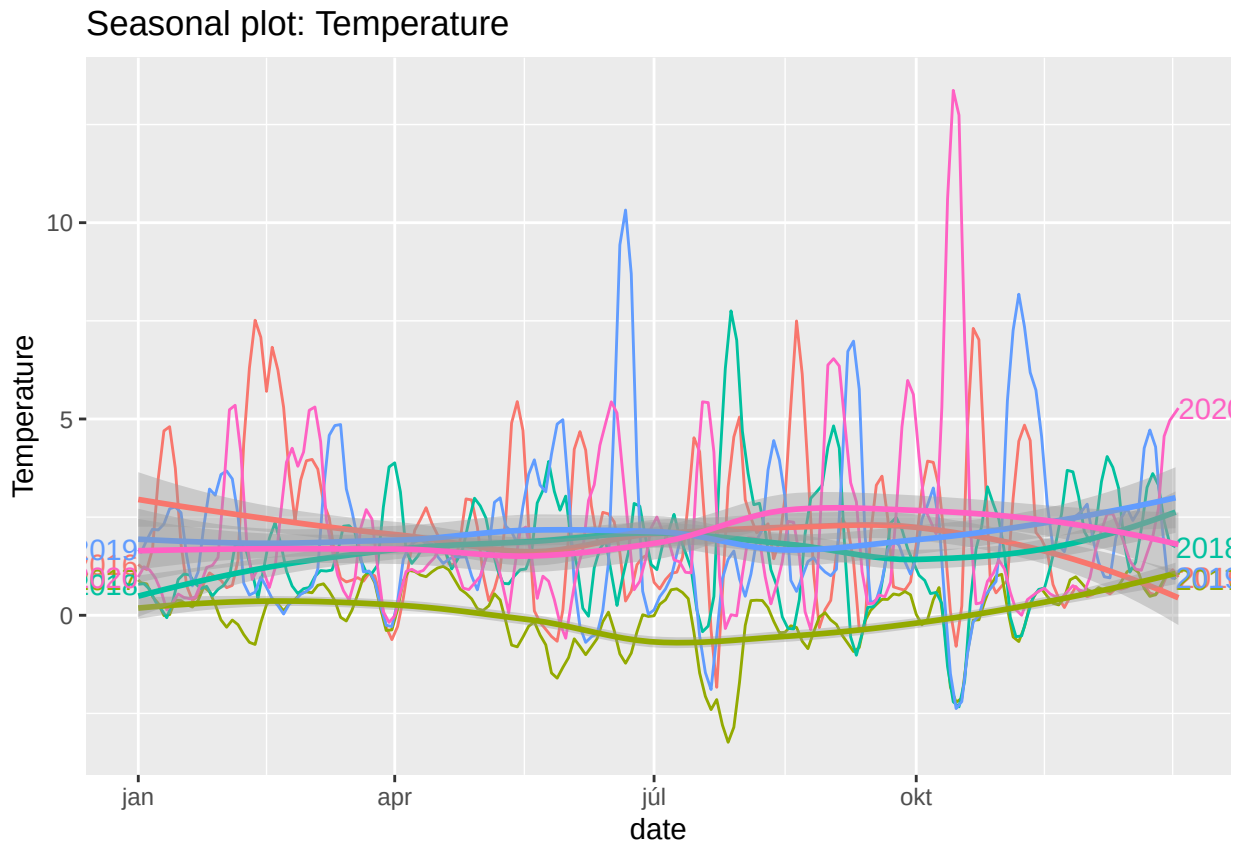
Môžeme vidieť, že sezónny príspevok sa v priebehu posledných rokov veľmi nemení.

```
m %>%
  components() %>%
  filter(year(date) > 2010) %>%
  gg_season(season_year, labels = "both") +
    labs(y = "Temperature",
       title = "Seasonal plot: Temperature")
```

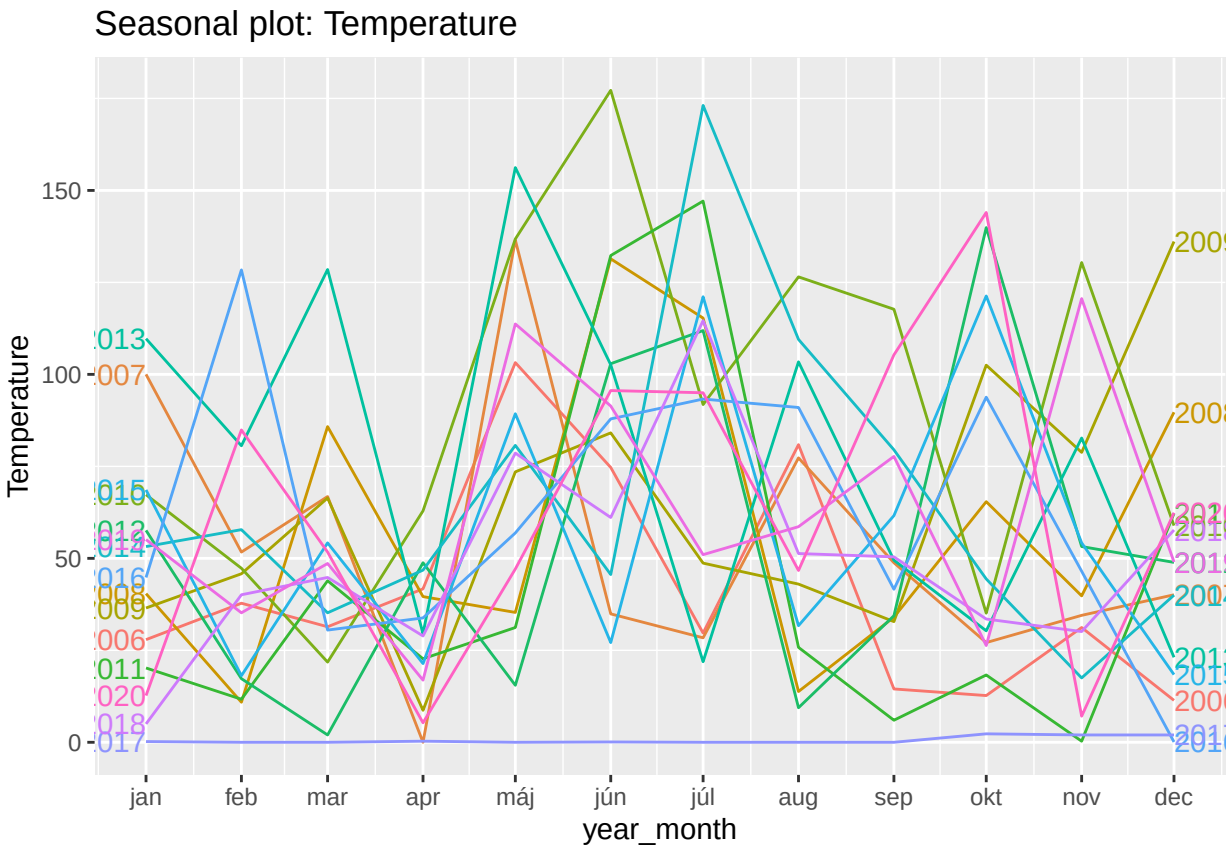



```
m %>%
  components() %>%
  filter(year(date) > 2015) %>%
  gg_season(trend, labels = "both") +
    geom_smooth() +
    labs(y = "Temperature",
       title = "Seasonal plot: Temperature")
```

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

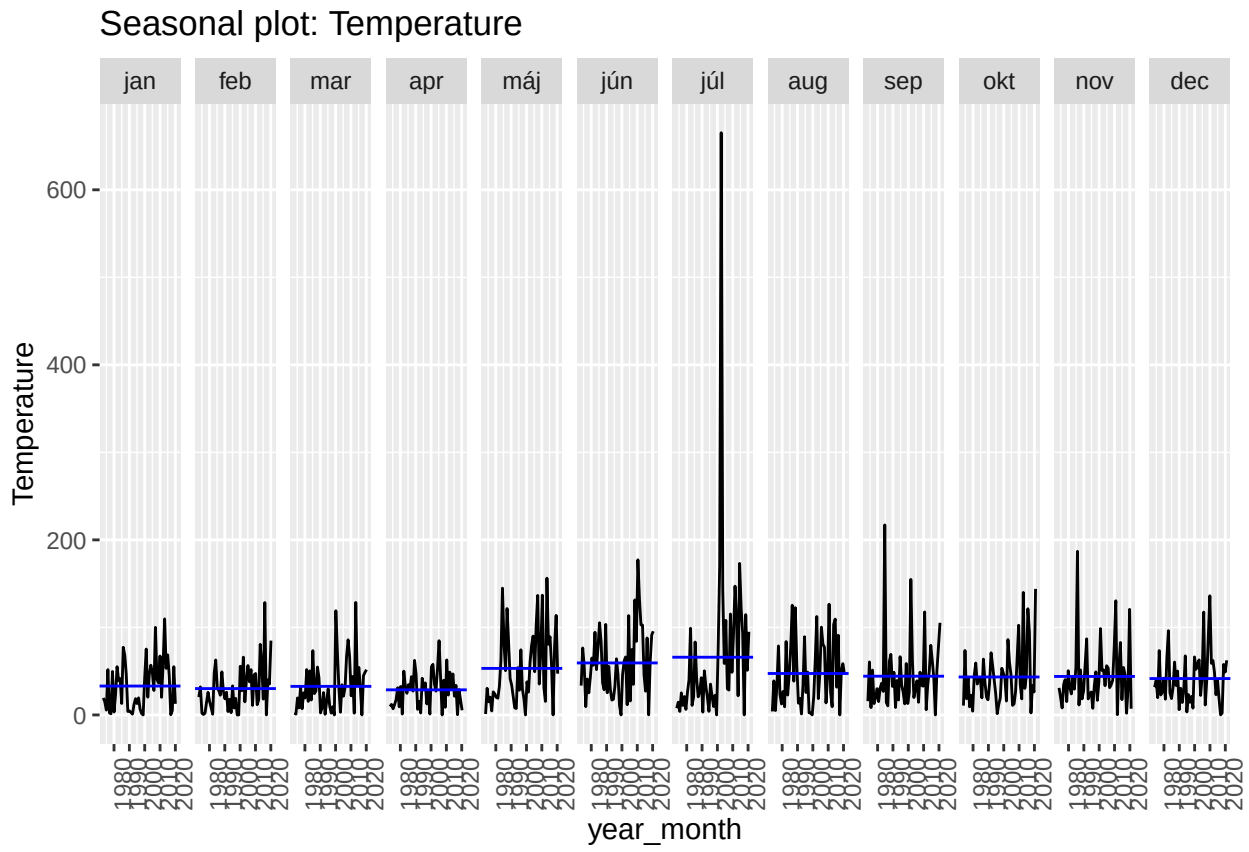



```
lp_df %>%
  as.data.frame() %>%
  dplyr::mutate(
    year_month = yearmonth(date)
  ) %>%
    dplyr::group_by(year_month) %>%
  dplyr::select(-date) %>%
  dplyr::summarise(LP_SUM = na.omit(sum(LP))) %>%
  as.data.frame() %>%
  distinct(year_month, .keep_all = TRUE) %>%
  as_tsibble(
    index = year_month
    ) %>%
  tsibble::fill_gaps() %>%
  dplyr::filter(year(year_month)>2005) %>%
  gg_season(LP_SUM, labels = "both") +
    labs(y = "Temperature",
       title = "Seasonal plot: Temperature")
```



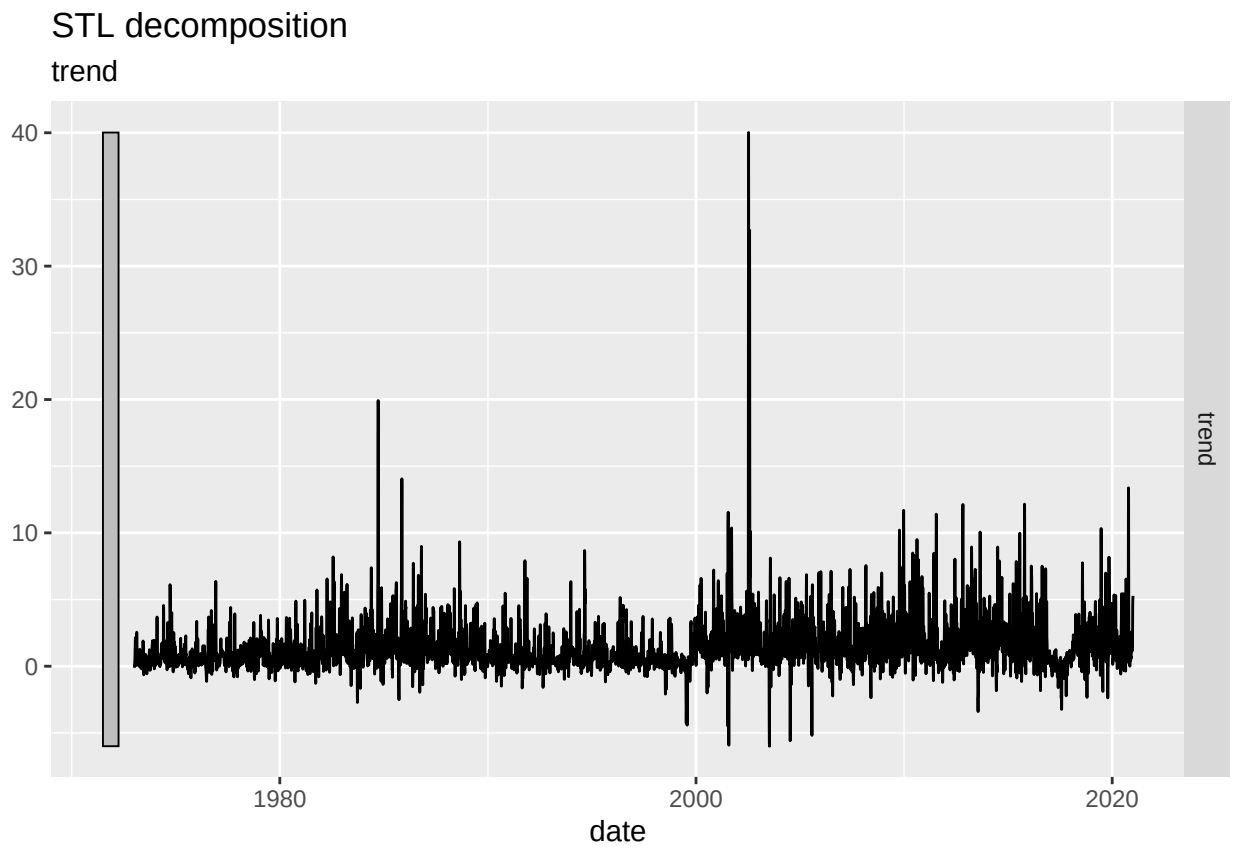
Vidíme ale ako vyzerajú jednotlivé mesiace. Napríklad máj, jún a júl majú výrazne väčšie priemery. TO isté je vidno aj na ostatných grafoch a prejavuje sa to v celom sezónnom príspevku.

```
lp_df %>%
  as.data.frame() %>%
  dplyr::mutate(
    year_month = yearmonth(date)
  ) %>%
  dplyr::group_by(year_month) %>%
  dplyr::select(-date) %>%
  dplyr::summarise(LP_SUM = na.omit(sum(LP))) %>%
  as.data.frame() %>%
  distinct(year_month, .keep_all = TRUE) %>%
  as_tsibble(
    index = year_month
    ) %>%
  tsibble::fill_gaps() %>%
  #dplyr::filter(year(year_month)>2010) %>%
  gg_subseries(LP_SUM, period = "1 year") +
  labs(y = "Temperature",
       title = "Seasonal plot: Temperature")
```



Príspevok trendu Trend na prvý pohlad nevyzerá rastúco ani klesajúco.

```
m %>%
  components() %>%
  autoplot(trend)
```



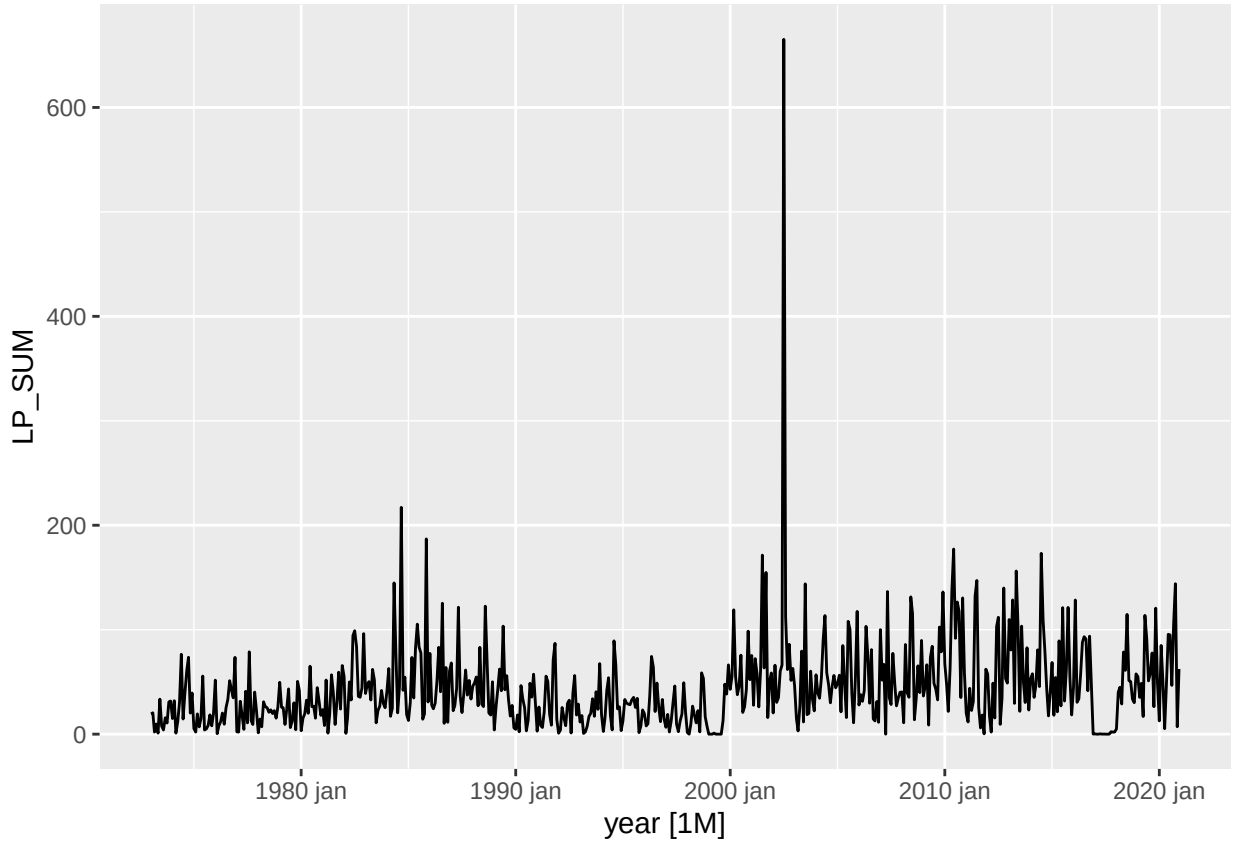
Môžeme sa ale pokúsiť predpovedať sklon trendu v budúcnosti pomocou lineránej regresie.

```
lp_df %>%
  as.data.frame() %>%
  dplyr::mutate(
    year = yearmonth(date)
  ) %>%
  dplyr::group_by(year) %>%
  dplyr::summarise(LP_SUM = na.omit(sum(LP))) %>%
  as.data.frame() %>%
  distinct(year, .keep_all = TRUE) %>%
  as_tsibble(
    index = year
    ) -> monthly_lp_df
```

Skúsime to najprv na mesačných súhrnoch zrážok.

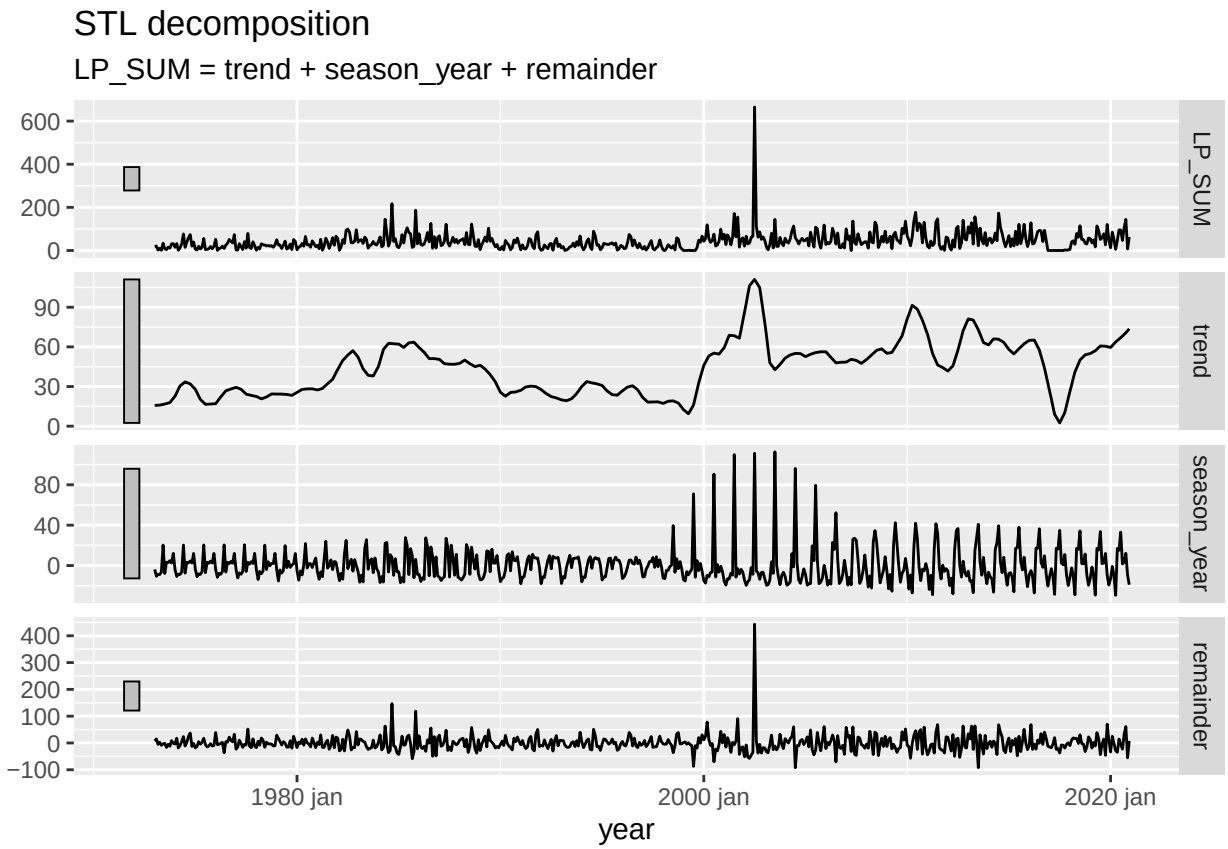
```
monthly_lp_df %>%
  autoplot()
```

```
## Plot variable not specified, automatically selected `.vars = LP_SUM`
```



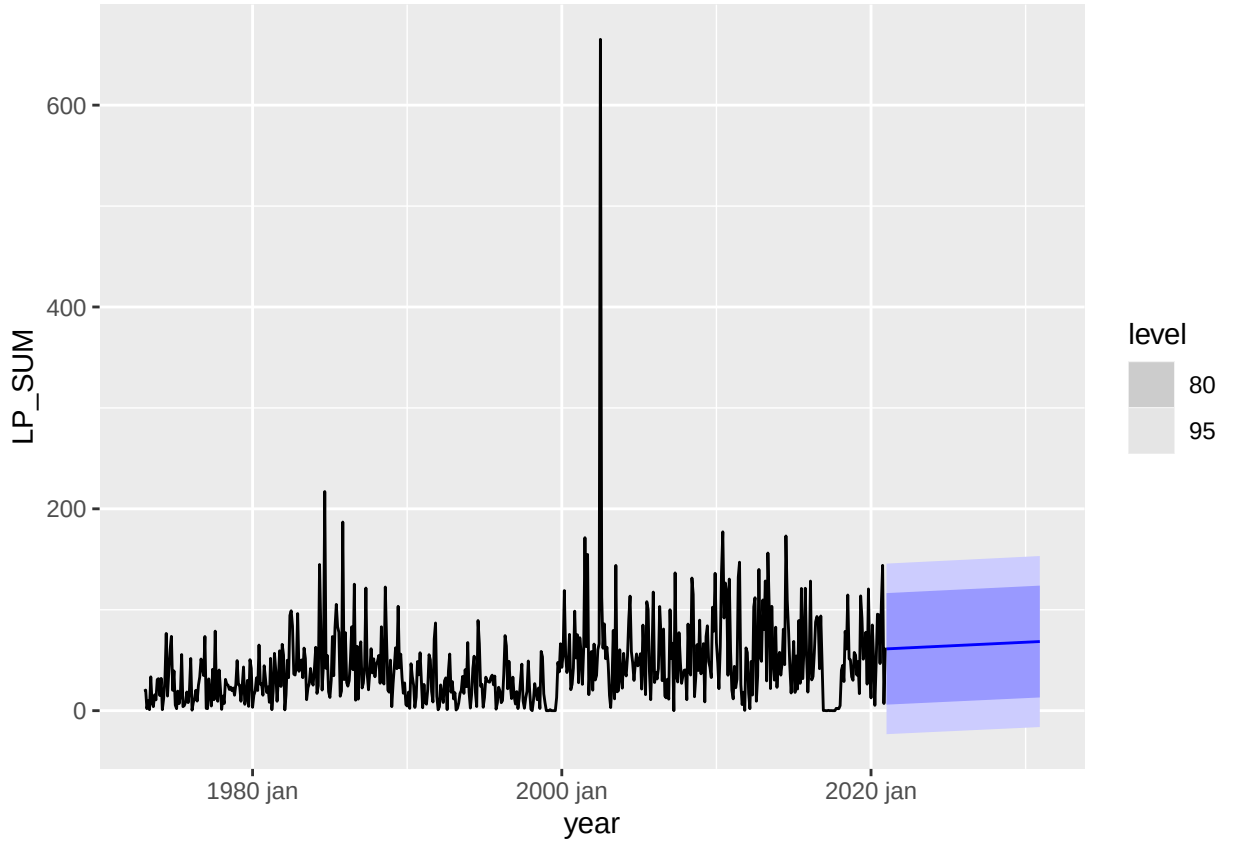
Sezonny prispevok je tam vidno ako ovela krašiu osciláciu a aj trend je oveľa jemnejší.

```
m <- monthly_lp_df %>%
   model(STL(LP_SUM ))

m %>%
  components() %>%
  autoplot()
```



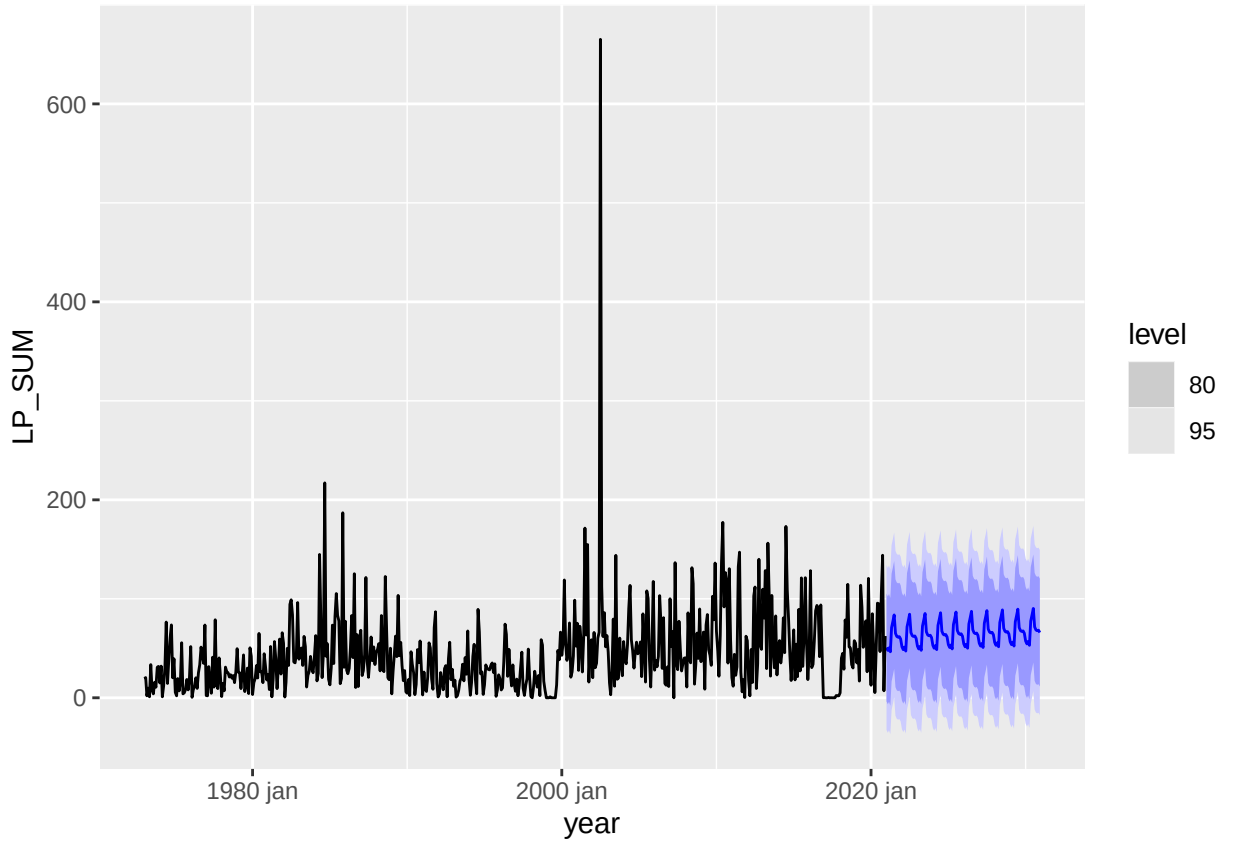
Natrénujeme model na predikciu trendu

```
monthly_lp_df %>%
  model(trend_model = TSLM(LP_SUM ~ trend())) -> m

m %>%
  forecast(h = "10 years") %>%
  autoplot(monthly_lp_df)
```



Aj model ktorý sa pokúsi predikovať aj sezónny príspevok.

```
monthly_lp_df %>%
  model(trend_model = TSLM(LP_SUM ~ trend() + season())) -> s_m
s_m %>%
  forecast(h = "10 years") %>%
  autoplot(monthly_lp_df)
```



Oba modeli ukazujú rastúcu tendenciu trendu a celkového množstva zrážok a potvrdzujú to aj vlastnosti danej lineárnej regresie.

```
report(m)
```

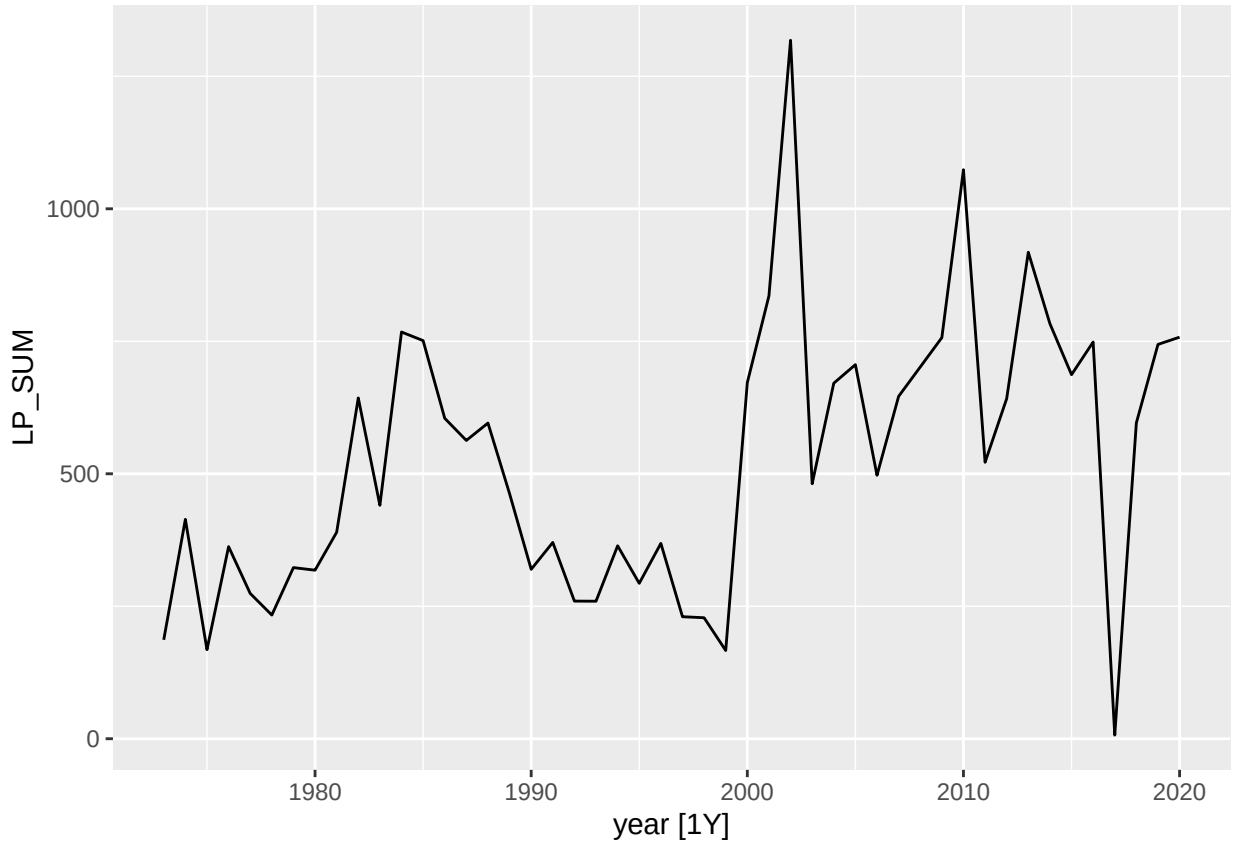
```
## Series: LP_SUM
## Model: TSLM
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.786 -23.481  -8.338  14.160 617.525
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.00299    3.58512   7.253 1.33e-12 ***
## trend()      0.06105    0.01077   5.670 2.26e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.97 on 574 degrees of freedom
## Multiple R-squared: 0.05304, Adjusted R-squared: 0.05139
## F-statistic: 32.15 on 1 and 574 DF, p-value: 2.2624e-08
```

Skúsime to urobiť aj pre ročný súhrn zrážok.

```
lp_df %>%
  as.data.frame() %>%
  dplyr::mutate(
```

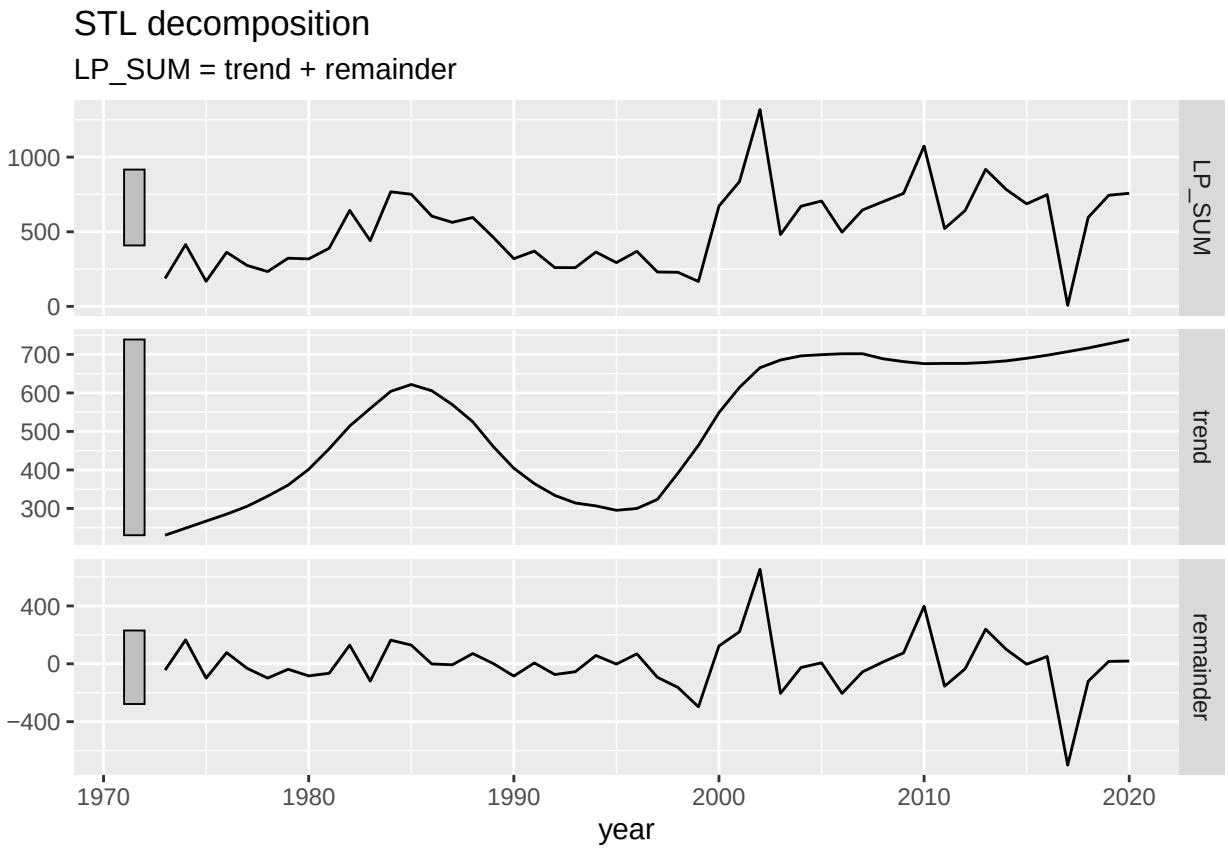
```
    year = year(date)
  ) %>%
  dplyr::group_by(year) %>%
  dplyr::summarise(LP_SUM = na.omit(sum(LP))) %>%
  as.data.frame() %>%
  distinct(year, .keep_all = TRUE) %>%
  as_tsibble(
    index = year
    ) -> yearly_lp_df

yearly_lp_df %>%
  autoplot()
```

```
## Plot variable not specified, automatically selected `.vars = LP_SUM`
```

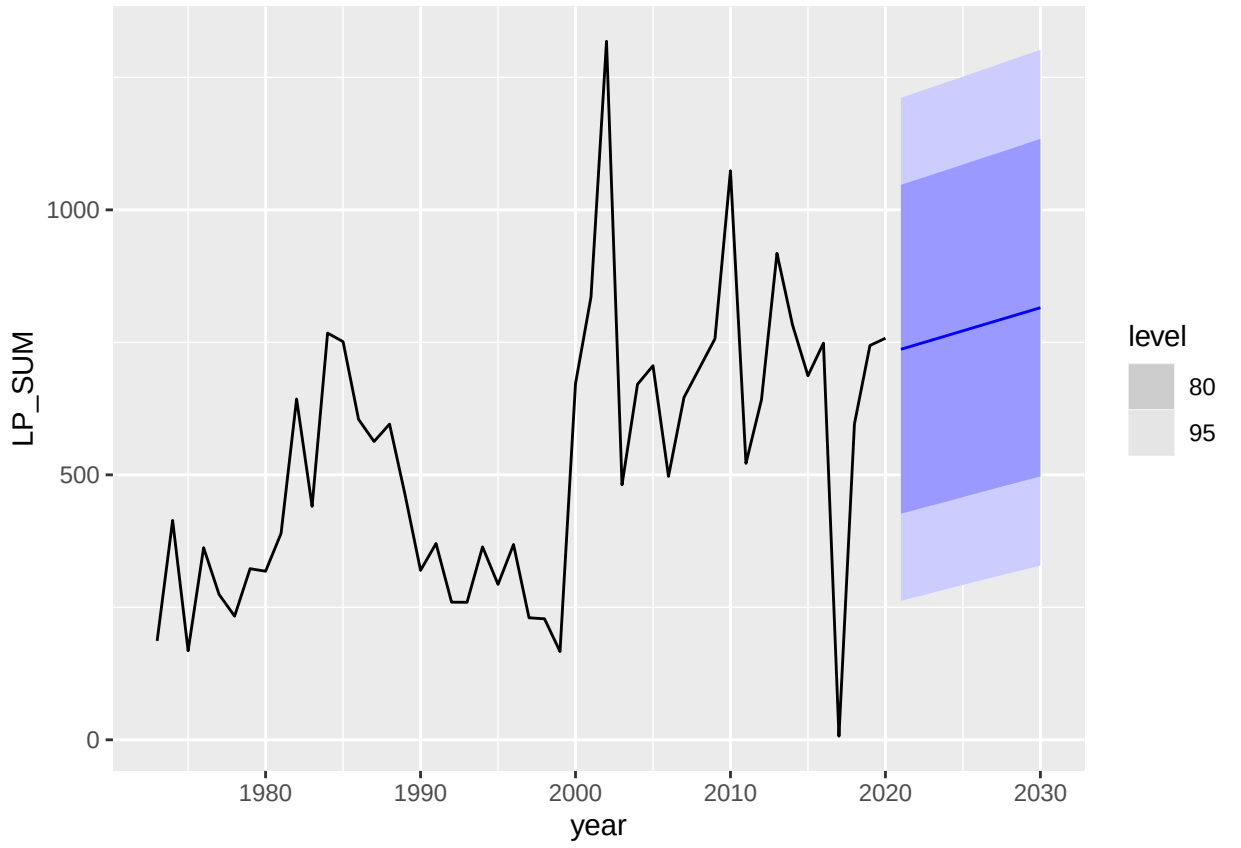


Trend je tu oveľa jenoduchší a sezónny príspevok chýba nakoľko sme už mimo sezón.

```
m <- yearly_lp_df %>%
   model(STL(LP_SUM ))

m %>%
  components() %>%
  autoplot()
```




```
yearly_lp_df %>%
  model(trend_model = TSLM(LP_SUM ~ trend())) -> m

m %>%
  forecast(h = "10 years") %>%
  autoplot(yearly_lp_df)
```



Vzniknutý model ukazuje rasť ročného súhrnu zrážok, čo potvrdzujú aj jeho vlastnosti. Je to štatisticky významne.

```
report(m)
```

```
## Series: LP_SUM
## Model: TSLM
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -695.168 -129.511    3.427  110.873  746.677
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  309.833     68.140   4.547 3.95e-05 ***
## trend()        8.716      2.421   3.600 0.000776 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 232.4 on 46 degrees of freedom
## Multiple R-squared: 0.2198,  Adjusted R-squared: 0.2029
## F-statistic: 12.96 on 1 and 46 DF, p-value: 0.00077557
```