

2.hypotéza

Denisa Mensatorisová a Adam Štuller

Druhá hypotéza, ktorú chceme overiť sa týka čisto zrážok. Chceme overiť, či sa mení distribúcia zrážok. Hypotéza znie, že počet dní kedy prší sa znižuje ale objem zrážok na jeden deň rastie. Predpokladáme teda, že počet suchých dní bez dažda rastie. Na iných miestach sme sa už zrážkam podrobnejšie venovali a preto pojdeme priamo na klasifikovanie dní.

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

Klasifikovanie dní podľa množstva zrážok

Jednodlivo dni klasifikujeme do šiestich kategórií podľa sumy všetkých zrážok počas dňa.

Kategória	Interval
0 - Ziadne dažď	<0>
1 - Jemný dažď	(0,10)
2 - Stredný dažď	<10, 35.5)
3 - Skor veľký dažď	<35.5, 64.4)
4 - Veľký dažď	<64.4, 124.4)
5 - Veľmi veľký dažď	124.4...

```
classify_rainy_day <- function (daily_rain) {  
  if(daily_rain == 0){  
    return (0)  
  }  
  else if( daily_rain < 10.0){  
    return (1)  
  }  
  else if( daily_rain < 35.5){  
    return (2)  
  }  
  else if( daily_rain < 64.4){  
    return (3)  
  }  
  else if( daily_rain < 124.4){  
    return (4)  
  }  
  else {  
    return (5)  
  }  
}
```

Túto klasifikáciu použijeme na vytvorenie novej datovej množiny.

```
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

lp_df %>%
  dplyr::mutate(
    day_class = map_dbl(LP, classify_rainy_day)
  ) -> lp_dc_df

lp_dc_df %>%
  autoplot( day_class) +
    labs(title = "Časový graf dní klasifikovaný podľa množstva zrážok",
         y = "Množstvo zrážok v mm",
         x = "Dátum"
    )

```



Vývoj početností tried v čase

Vypočítame počty rôznych tried dní v mesiacoch a pozrieme si vývoj početností jednotlivých tried v čase.

```

lp_dc_df %>%
  as.data.frame()%>%
  dplyr::mutate(
    year = year(date)
  ) %>%
  dplyr::group_by(year, day_class) %>%
  dplyr::summarise(
    n_day_class = n()
  )

```

```
) %>%
as.data.frame() -> class_n_df
```

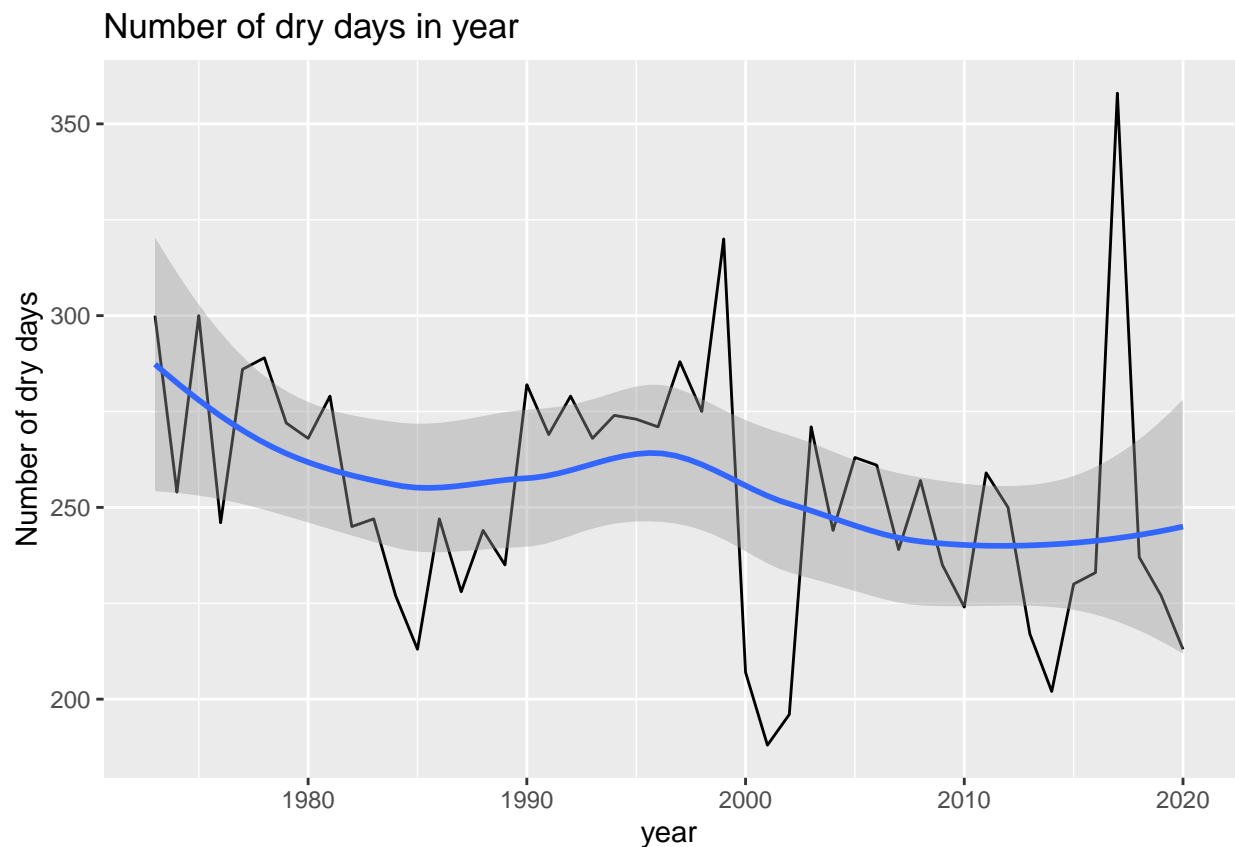
`summarise()` has grouped output by 'year'. You can override using the `.groups` argument.

Suché dni

Hned tu môžeme vidieť, že počet suchých dni prekvapivo vyzerá akoby klesal.

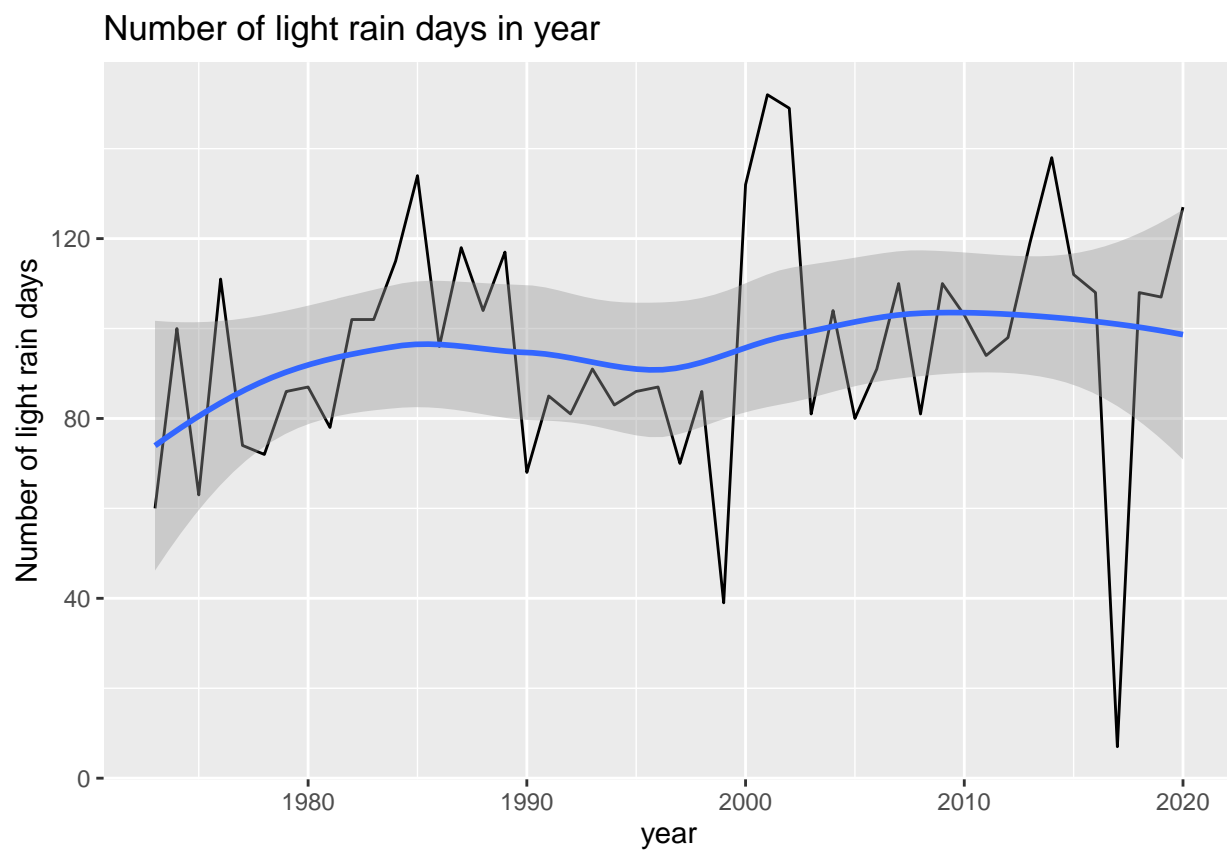
```
class_n_df %>%
  filter(day_class == 0) %>%
  as_tsibble(
    index = year
  ) %>%
  autoplot(n_day_class) +
  geom_smooth() +
  labs(title = "Number of dry days in year",
       y = "Number of dry days",
       x = "year"
  )
```

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



Slabý dážď

```
class_n_df %>%  
  filter(day_class == 1) %>%  
  as_tsibble(  
    index = year  
  ) %>%  
  autoplot(n_day_class) +  
    geom_smooth() +  
    labs(title = "Number of light rain days in year",  
         y = "Number of light rain days",  
         x = "year"  
    )  
  
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



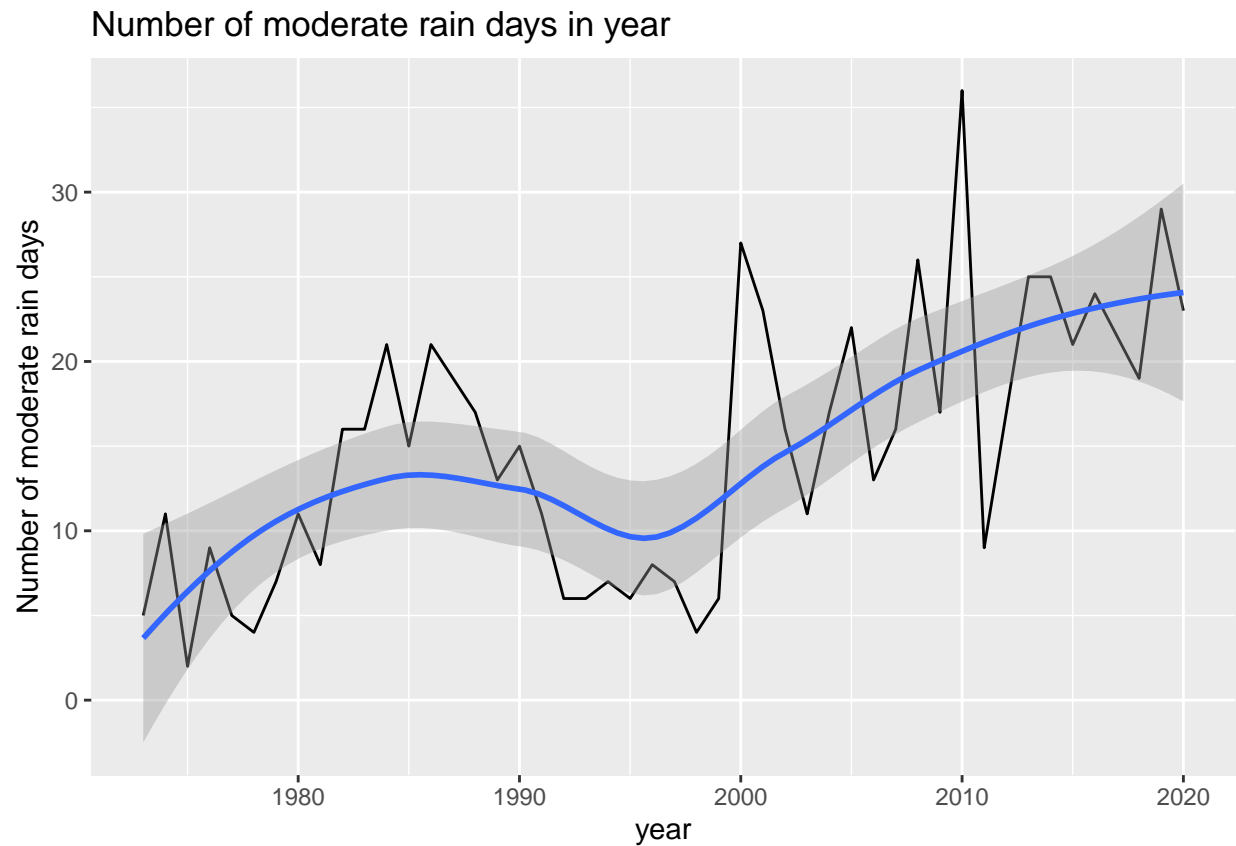
Stredný dážď

Počet dní so stredným daždom prekvapivo rastie.

```
class_n_df %>%  
  filter(day_class == 2) %>%  
  as_tsibble(  
    index = year  
  ) %>%  
  autoplot(n_day_class) +  
    geom_smooth() +
```

```
labs(title = "Number of moderate rain days in year",
      y = "Number of moderate rain days",
      x = "year"
    )
```

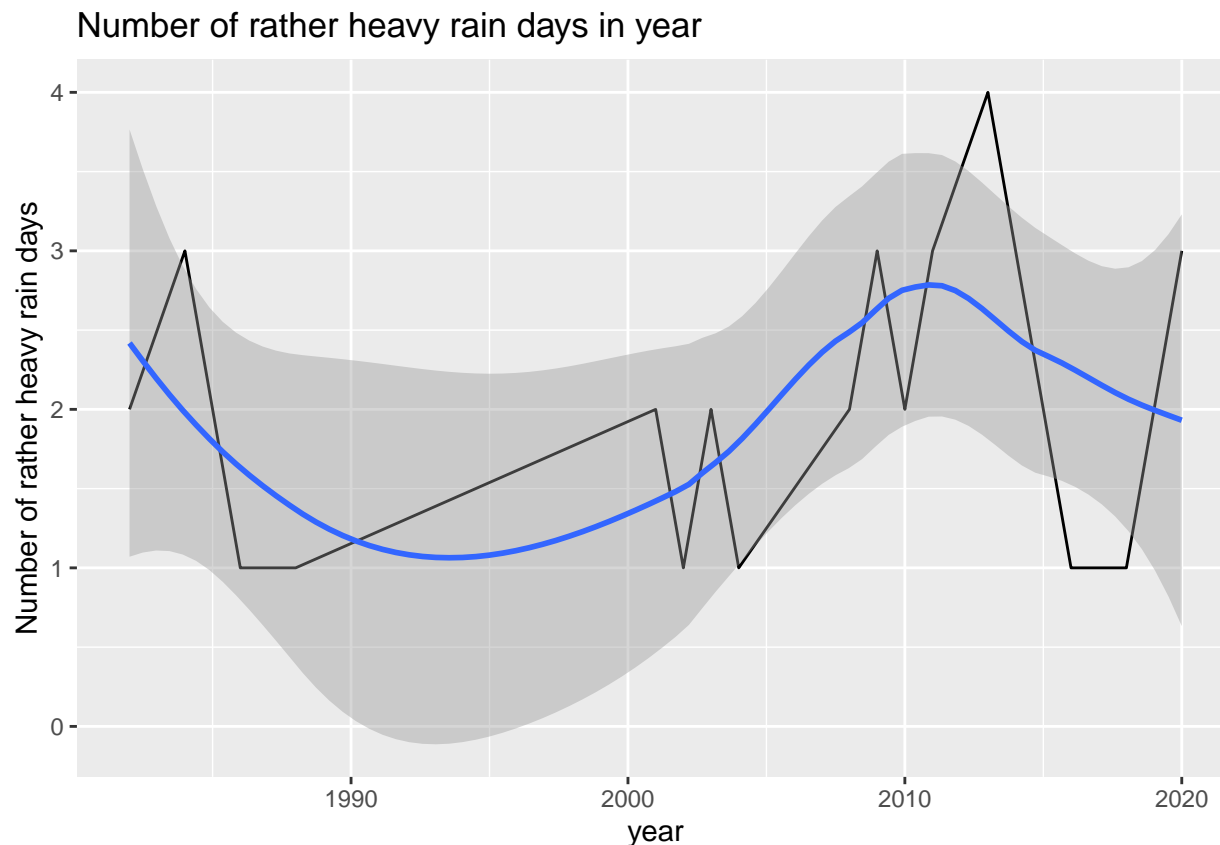
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Skôr silný dážď

```
class_n_df %>%
  filter(day_class == 3) %>%
  as_tsibble(
    index = year
  ) %>%
  autoplot(n_day_class) +
  geom_smooth() +
  labs(title = "Number of rather heavy rain days in year",
        y = "Number of rather heavy rain days",
        x = "year"
      )
```

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



Ostatné triedy sa neoplatilo ani zobrazovať, pretože ich príkladov bolo príliš málo.

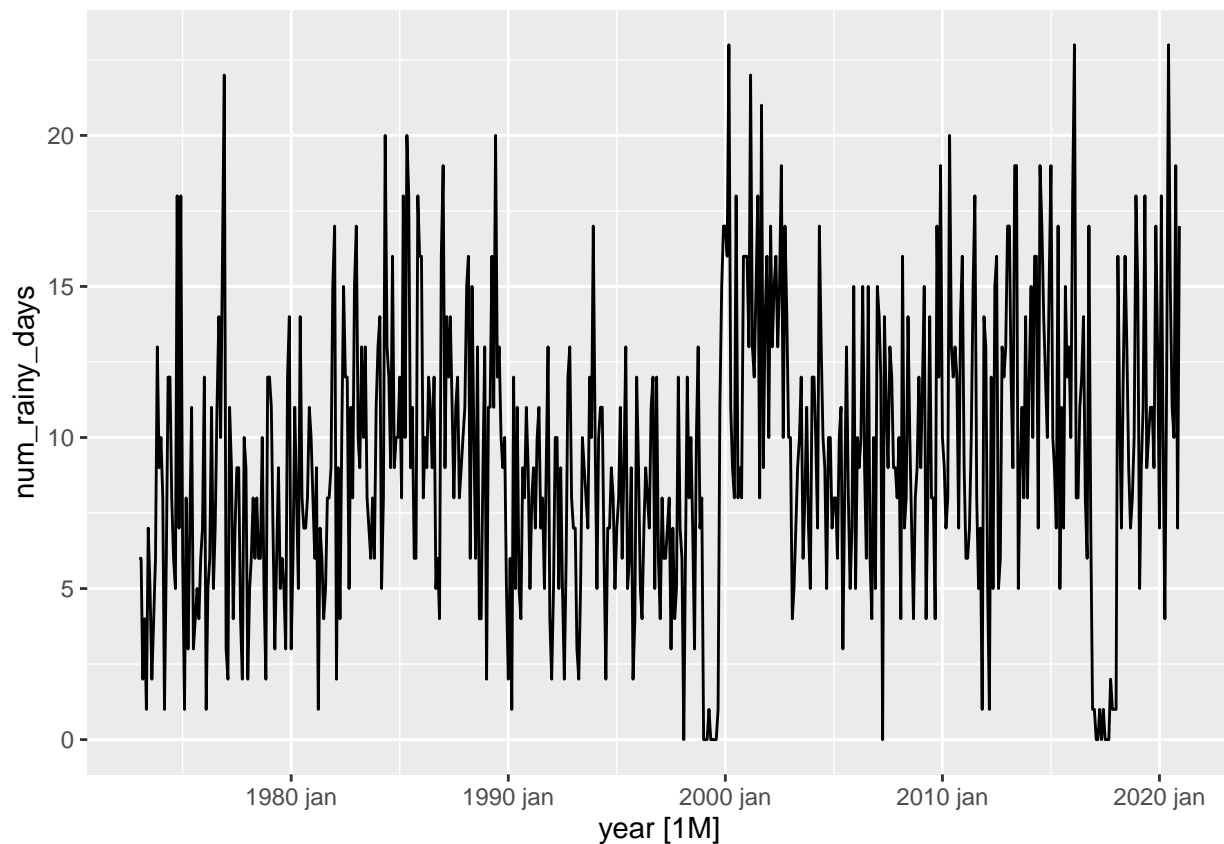
Overenie hypotézy

Vyvoríme tri stĺpce. Jeden bude obsahovať všetky dni počty dní bez dažda v mesiaci. Ďalší počet všetkých dní kedy pršalo aspoň trochu. Posledný stĺpec bude obsahovať množstvo zrážok v mesiaci pripadajúce na jeden deň - teda pomer množstvo zrážok lomeno počet dní kedy pršalo.

```
lp_dc_df %>%
  as.data.frame()%>%
  dplyr::mutate(
    year = yearmonth(date)
  ) %>%
  dplyr::group_by(year) %>%
  dplyr::summarise(
    num_dry_days = sum(ifelse(day_class == 0, 1, 0)),
    num_rainy_days = sum(ifelse(day_class == 0, 0, 1)),
    sum_rain = sum(LP)
  ) %>%
  mutate(
    precipitation_ratio = num_rainy_days / sum_rain
  ) %>%
  as.data.frame() %>%
  as_tsibble(
    index = year
  ) -> df
```

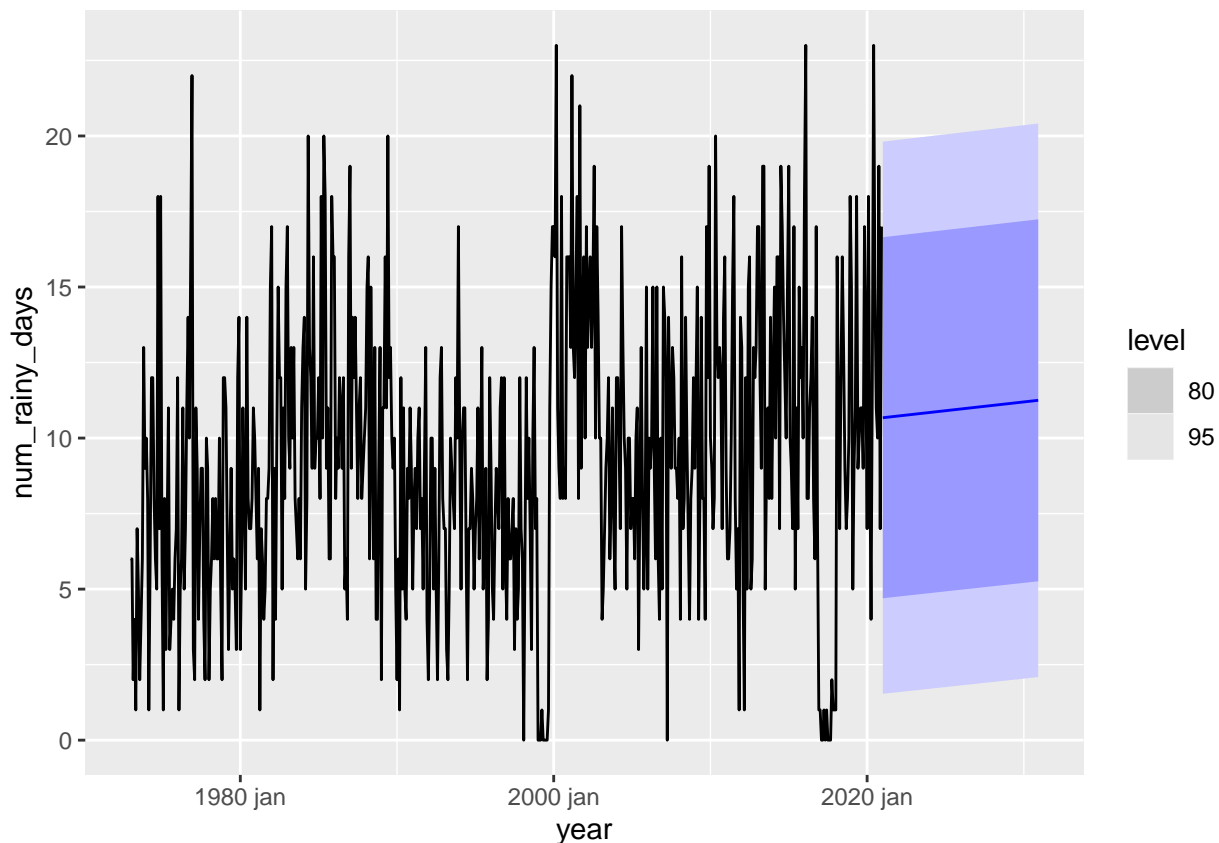
Vývoj počtu dažďových dní v mesiaci

```
df %>%  
  autoplot(num_rainy_days)
```



Vydime, že už na prvý pohľad vyzerá, že množstvo dažďových dní rastie a nie klesá.

```
df %>%  
  model(trend_model = TSLM(num_rainy_days ~ trend())) -> m  
  
m %>%  
  forecast(h = "10 years") %>%  
  autoplot(df)
```

Po použití lineárnej regresie vidíme, že pri predikcii na 10 rokov dopredu je vzniknutá priamka rastúca celkom výrazne a pri pohľade na opis modelu vidíme, že je to aj štatisticky signifikantne. Vyvodzujeme teda preto, že z našich dát vyplíva, že počet dní kedy prší rastie namiesto toho aby klesal.

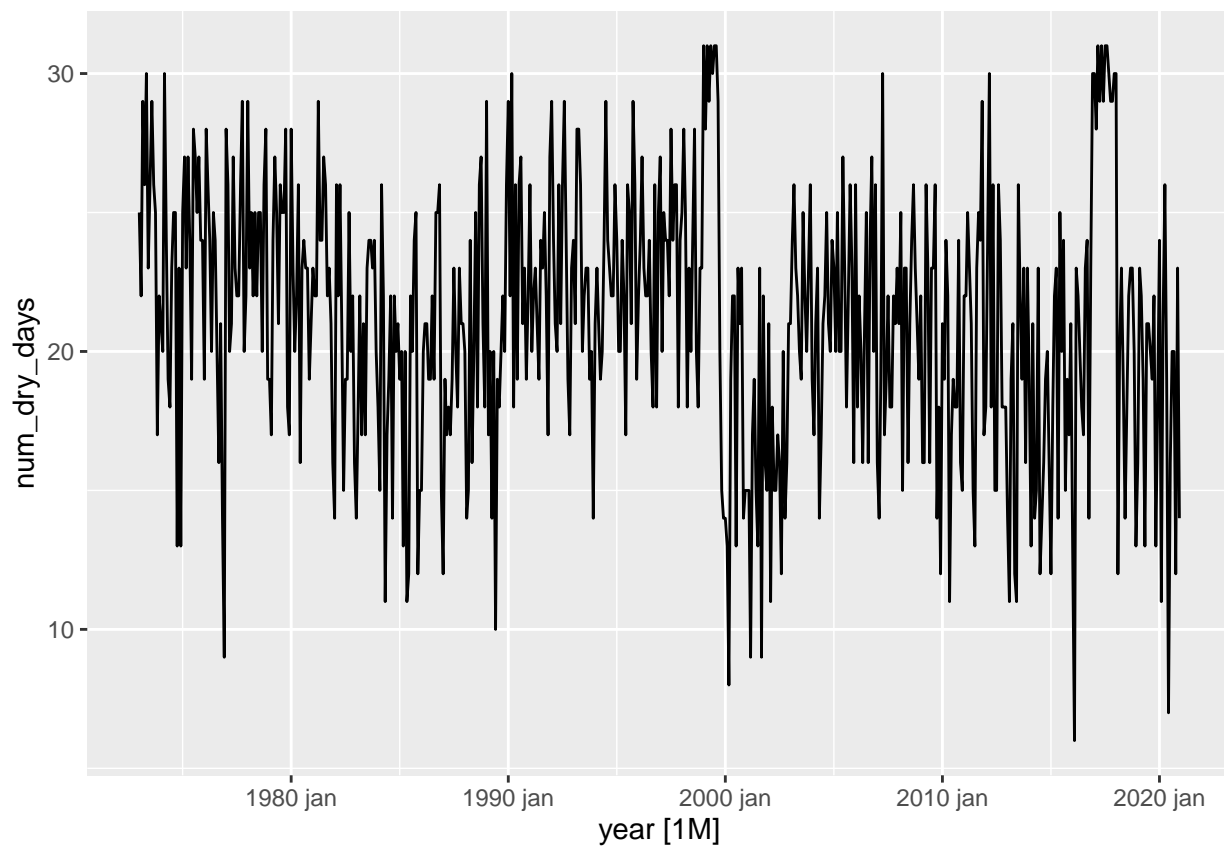
```
report(m)
```

```
## Series: num_rainy_days
## Model: TSLM
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4790  -3.0859  -0.2567   2.9156  13.8916
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.875682   0.387709  20.313  < 2e-16 ***
## trend()      0.004848   0.001164   4.164 3.61e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.646 on 574 degrees of freedom
## Multiple R-squared:  0.02932, Adjusted R-squared:  0.02763
## F-statistic: 17.34 on 1 and 574 DF, p-value: 3.6129e-05
```

Vývoj počtu suchých dní v mesiaci

Postupujeme rovnako ako v prvom prípade.

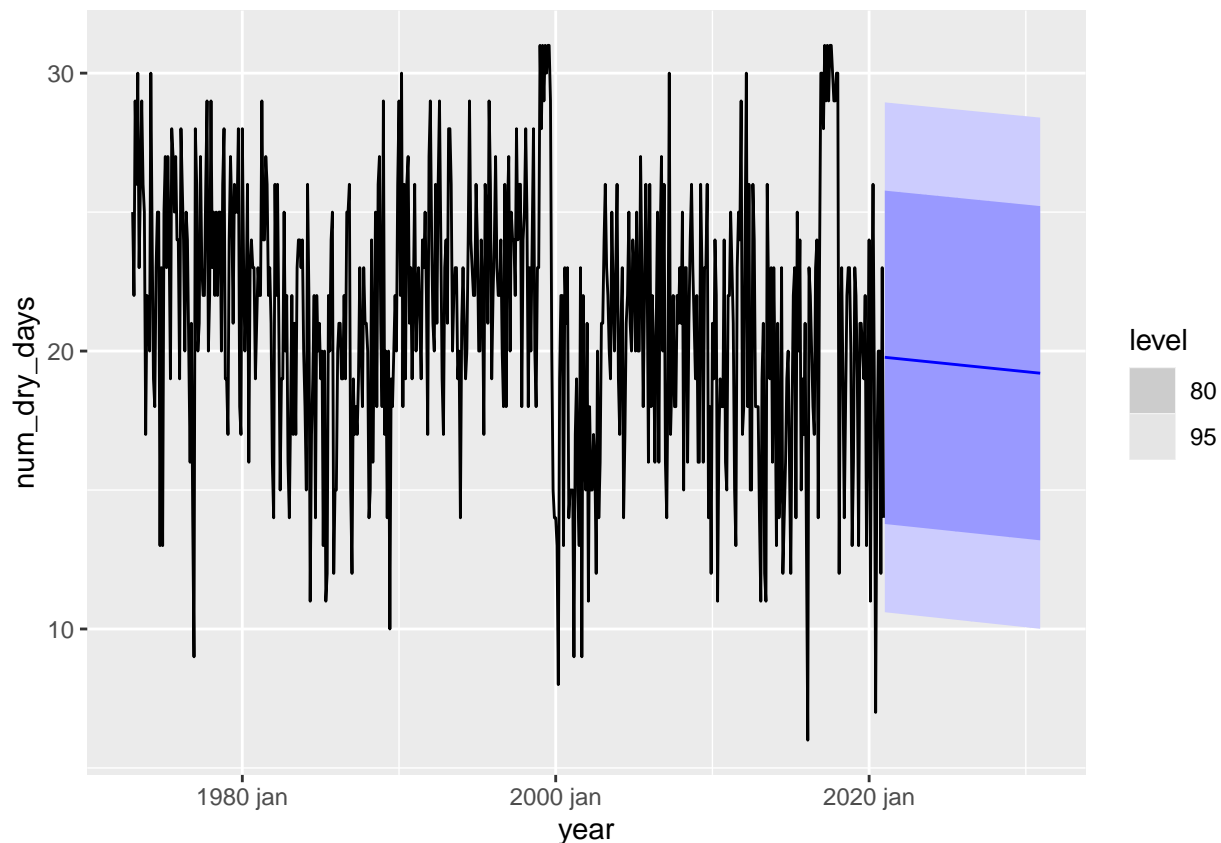
```
df %>%
  autoplot(num_dry_days)
```



Vidíme, že tento graf vyzerá skôr klesajúco.

```
df %>%
  model(trend_model = TSLM(num_dry_days ~ trend())) -> m

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



Lineárna regresia nám to iba potvrdzuje a model hovorí, že je to signifikantné a hodnota t je záporná. (-4.115).

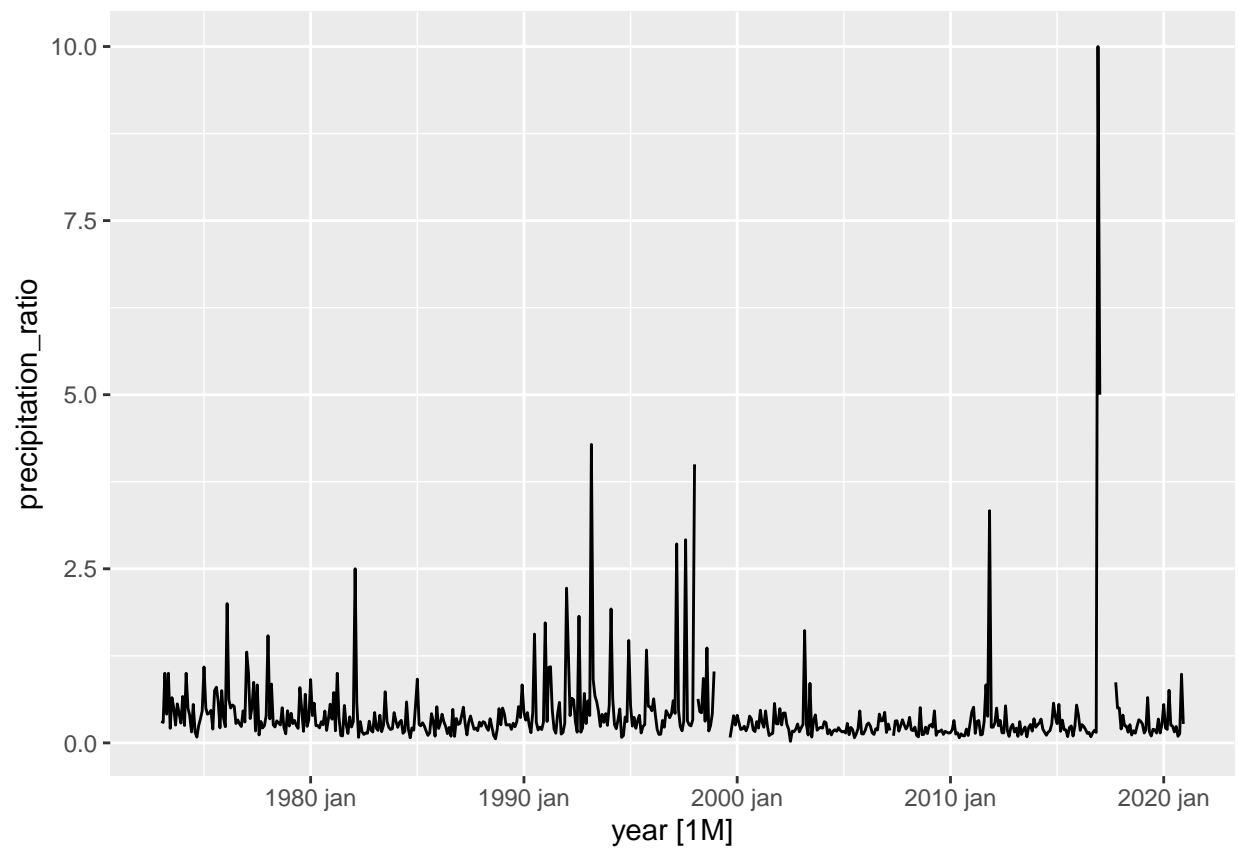
```
report(m)
```

```
## Series: num_dry_days
## Model: TSLM
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.0595  -2.9098   0.2191   2.9307  11.0271
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.550622   0.389145  57.949  < 2e-16 ***
## trend()     -0.004809   0.001169  -4.115 4.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.664 on 574 degrees of freedom
## Multiple R-squared:  0.02866, Adjusted R-squared:  0.02696
## F-statistic: 16.93 on 1 and 574 DF, p-value: 4.4373e-05
```

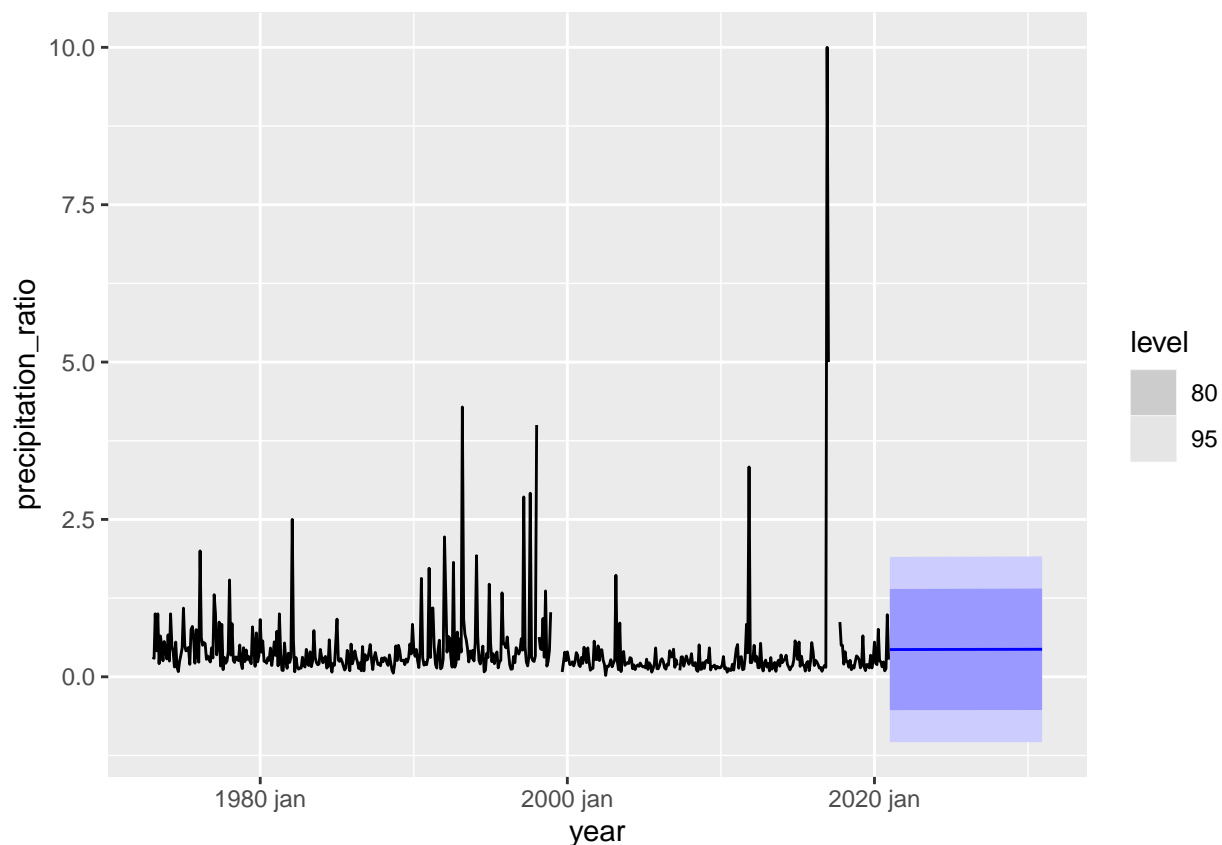
Vývoj počtu dažďových dní v mesiaci

POzrieme sa aj na to ako sa vyvíja množstvo zrážok pripadajúce na jeden deň v mesiaci.

```
df %>%
  autoplot(precipitation_ratio)
```



```
df %>%  
  model(trend_model = TSLM(precipitation_ratio ~ trend())) -> m  
  
m %>%  
  forecast(h = "10 years") %>%  
  autoplot(df)
```



```
report(m)
```

```
## Series: precipitation_ratio
## Model: TSLM
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.402731 -0.242481 -0.160051 -0.000495  9.568601
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.127e-01  6.267e-02   6.585 1.05e-10 ***
## trend()      3.537e-05  1.898e-04   0.186   0.852
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.748 on 559 degrees of freedom
## Multiple R-squared:  6.213e-05,    Adjusted R-squared: -0.001727
## F-statistic: 0.03474 on 1 and 559 DF, p-value: 0.85222
```

Vidíme teda, že tento pomer sa z dlhodobého hľadiska štatisticky významne nemení.

Zhodnotenie

Overovali sme hypotézu, že distribúcia zrážok sa mení. Zistili sme však, že sa to nedeje ako sme si najprv mysleli. Z našich dát vypláva, že množstvo zrážok rastie, rastie aj množstvo dní kedy prší. Naopak klesá počet suchých dní kedy neprší vôbec. množstvo zrážok pripadajúcich na jeden deň v mesiaci sa ale nemení -

zrážok je celkovo viac.