Institut za matematiku i informatiku
Prirodno-matematički fakultet
Univerzitet u Kragujevcu

Seminarski rad iz predmeta
Predstavljanje i tumačenje podataka

Eksploratorna analiza

Globalne temperature od 1775. godine

Student:
Nikola Ratinac 1030/2019

Profesor:
dr Marinko Timotijević

Februar 2021.

Sadržaj

Predmet rada jeste utvrđivanje prosečnih mesečnih temperatura po godini i zemlji i utrvrđivanje globalnog trenda prosečne godišnje temperature.

Baza podataka

Baza podataka se sastoji od 6 .csv fajlova.

```
system("ls input", intern=TRUE)
```

```
## [1] "EmissionData.csv"
## [2] "GlobalLandTemperaturesByCity.csv"
## [3] "GlobalLandTemperaturesByCountry.csv"
## [4] "GlobalLandTemperaturesByMajorCity.csv"
## [5] "GlobalLandTemperaturesByState.csv"
## [6] "GlobalTemperatures.csv"
```

Učitaćemo podatke i ispitati njihovu strukturu.

```
df_city <- fread("input/GlobalLandTemperaturesByCity.csv")
df_country <- read.csv("input/GlobalLandTemperaturesByCountry.csv", stringsAsFactors = F)
df_emission <- read.csv("input/EmissionData.csv", check.names = FALSE, stringsAsFactors = F, header = T
df_global <- read.csv("input/GlobalTemperatures.csv", stringsAsFactors = F)
```

```
str(df_city)
```

```
## Classes 'data.table' and 'data.frame':   8599212 obs. of  7 variables:
##  $ dt                           : IDate, format: "1743-11-01" "1743-12-01" ...
##  $ AverageTemperature           : num  6.07 NA NA NA NA ...
##  $ AverageTemperatureUncertainty: num  1.74 NA NA NA NA ...
##  $ City                         : chr  "Århus" "Århus" "Århus" "Århus" ...
##  $ Country                      : chr  "Denmark" "Denmark" "Denmark" "Denmark" ...
##  $ Latitude                     : chr  "57.05N" "57.05N" "57.05N" "57.05N" ...
##  $ Longitude                    : chr  "10.33E" "10.33E" "10.33E" "10.33E" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

Podaci koji se odnose na temperature po gradovima su podeljeni po sledećim kolonama:

- dt - datum obzervacije izvedena na mesečnom nivou
- AverageTemperature - prosečna temperatura tog meseca
- AverageTemperatureUncertainty - nesigurnost u podatak o prosečnoj temperaturi
- City - naziv grada
- Country - naziv države
- Latitude - geografska širina
- Longitude - geografska dužina

```
summary(df_city)
```

```
##        dt             AverageTemperature AverageTemperatureUncertainty
##  Min.   :1743-11-01   Min.   :-42.7      Min.   : 0.0
```

```
##  1st Qu.:1860-06-01   1st Qu.: 10.3      1st Qu.: 0.3
##  Median :1911-09-01   Median : 18.8      Median : 0.6
##  Mean   :1907-10-21   Mean   : 16.7      Mean   : 1.0
##  3rd Qu.:1962-09-01   3rd Qu.: 25.2      3rd Qu.: 1.3
##  Max.   :2013-09-01   Max.   : 39.7      Max.   :15.4
##                       NA's   :364130     NA's   :364130
##      City              Country            Latitude          Longitude
##  Length:8599212     Length:8599212     Length:8599212     Length:8599212
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

Pogledajmo od čega se sastoji skup podataka vezan za temperature po državama.

```
str(df_country)
```

```
## 'data.frame':    577462 obs. of  4 variables:
##  $ dt                           : chr  "1743-11-01" "1743-12-01" "1744-01-01" "1744-02-01" ...
##  $ AverageTemperature           : num  4.38 NA NA NA NA ...
##  $ AverageTemperatureUncertainty: num  2.29 NA NA NA NA ...
##  $ Country                      : chr  "Åland" "Åland" "Åland" "Åland" ...
```

Podaci koji se odnose na temperature po državama su podeljenji po sledećim kolonama:

- dt - datum obzervacije izvedena na mesečnom nivou
- AverageTemperature - prosečna temperatura tog meseca
- AverageTemperatureUncertainty - nesigurnost u podatak o prosečnoj temperaturi
- Country - naziv države

```
summary(df_country)
```

```
##       dt            AverageTemperature AverageTemperatureUncertainty
##  Length:577462      Min.   :-37.66     Min.   : 0.05
##  Class :character   1st Qu.: 10.03     1st Qu.: 0.32
##  Mode  :character   Median : 20.90     Median : 0.57
##                     Mean   : 17.19     Mean   : 1.02
##                     3rd Qu.: 25.81     3rd Qu.: 1.21
##                     Max.   : 38.84     Max.   :15.00
##                     NA's   :32651      NA's   :31912
##    Country
##  Length:577462
##  Class :character
##  Mode  :character
##
##
##
##
```

Predstavimo i podatke o emitovanju CO2 u atmosferu.

```r
str(df_emission[1:15])
```

```
## 'data.frame':    231 obs. of  15 variables:
##  $ Country: chr  "Afghanistan" "Africa" "Albania" "Algeria" ...
##  $ 1751   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ 1752   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ 1753   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ 1754   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ 1755   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ 1756   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ 1757   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ 1758   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ 1759   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ 1760   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ 1761   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ 1762   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ 1763   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ 1764   : int  0 0 0 0 0 0 0 0 0 0 ...
```

Ovaj skup podataka nije pogodan za dalju obradu ovakav kakav jeste, stoga ćemo njime podrobnije pozabaviti prilikom pripreme podataka.

```r
str(df_global)
```

```
## 'data.frame':    3192 obs. of  9 variables:
##  $ dt                                       : chr  "1750-01-01" "1750-02-01" "1750-03-01" "1750-04-0
##  $ LandAverageTemperature                   : num  3.03 3.08 5.63 8.49 11.57 ...
##  $ LandAverageTemperatureUncertainty        : num  3.57 3.7 3.08 2.45 2.07 ...
##  $ LandMaxTemperature                       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ LandMaxTemperatureUncertainty            : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ LandMinTemperature                       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ LandMinTemperatureUncertainty            : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ LandAndOceanAverageTemperature           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ LandAndOceanAverageTemperatureUncertainty: num  NA NA NA NA NA NA NA NA NA NA ...
```

```r
summary(df_global)
```

```
##       dt            LandAverageTemperature LandAverageTemperatureUncertainty
##  Length:3192        Min.   :-2.080         Min.   :0.0340
##  Class :character   1st Qu.: 4.312         1st Qu.:0.1867
##  Mode  :character   Median : 8.611         Median :0.3920
##                     Mean   : 8.375         Mean   :0.9385
##                     3rd Qu.:12.548         3rd Qu.:1.4192
##                     Max.   :19.021         Max.   :7.8800
##                     NA's   :12             NA's   :12
##  LandMaxTemperature LandMaxTemperatureUncertainty LandMinTemperature
##  Min.   : 5.90      Min.   :0.0440                Min.   :-5.407
##  1st Qu.:10.21      1st Qu.:0.1420                1st Qu.:-1.335
##  Median :14.76      Median :0.2520                Median : 2.950
##  Mean   :14.35      Mean   :0.4798                Mean   : 2.744
##  3rd Qu.:18.45      3rd Qu.:0.5390                3rd Qu.: 6.779
```

```
##  Max.   :21.32      Max.   :4.3730                Max.   : 9.715
##  NA's   :1200       NA's   :1200                  NA's   :1200
##  LandMinTemperatureUncertainty LandAndOceanAverageTemperature
##  Min.   :0.0450                Min.   :12.47
##  1st Qu.:0.1550                1st Qu.:14.05
##  Median :0.2790                Median :15.25
##  Mean   :0.4318                Mean   :15.21
##  3rd Qu.:0.4582                3rd Qu.:16.40
##  Max.   :3.4980                Max.   :17.61
##  NA's   :1200                  NA's   :1200
##  LandAndOceanAverageTemperatureUncertainty
##  Min.   :0.0420
##  1st Qu.:0.0630
##  Median :0.1220
##  Mean   :0.1285
##  3rd Qu.:0.1510
##  Max.   :0.4570
##  NA's   :1200
```

Priprema podataka

Pre nego sto pocnemo sa analizom podataka, potrebno je podtatke precistiti od nepostojecih vrednosti, ukloniti kolone koje nam ne govore nista i podatke transformisati na nacin pogodan za obradu i vizuelizaciju.

Transformacija podataka

U ovom odeljku se bavimo organizaciom podataka tako da oni imaju najvise smisla za onog koji ce se njima baviti. Posto u nekom trenutku treba implementirati modele masinskog ucenja koji su zasnovani na vremenskim serijama, potrebno je datume pretvoriti iz string reprezentacije u `Date` reprezentaciju koja je:

- Razumljiva R-u
- Pogodna za uspostavljanje hronoloskog poretka obzervacija

```
df_country$dt <- as.Date(df_country$dt)
df_city$dt <- as.Date(df_city$dt)
df_global$dt <- as.Date(df_global$dt)
```

Korisno je i imati zasebne vrednosti za mesece i godine.

```
df_country$year <- format(as.Date(df_country$dt), "%Y")
df_country$month <- format(as.Date(df_country$dt), "%m")
df_global$year <- format(as.Date(df_global$dt), "%Y")
df_global$month <- format(as.Date(df_global$dt), "%m")
df_city$year <- format(as.Date(df_city$dt), "%Y")
df_city$month <- format(as.Date(df_city$dt), "%m")
```

Pogodno podatke transformisati na nacin da budu u nekoj meri smisleni onome kome je zadatak da ih analizira. Stoga je korisno kolone nazivati smisleno i koncizno, a skupove podataka pretvoriti u one tipove koji su najpogodniji za analizu.

```
df_country <- as_tibble(df_country)
df_city <- as_tibble(df_city)
df_global <- as_tibble(df_global)

df_country <- df_country %>% rename(
    avgT = AverageTemperature,
    avgTU = AverageTemperatureUncertainty
)

df_city <- df_city %>% rename(
    avgT = AverageTemperature,
    avgTU = AverageTemperatureUncertainty,
    Lat = Latitude,
    Lng = Longitude
)

df_global <- df_global %>% rename(
    avgT = LandAverageTemperature,
    avgTU = LandAverageTemperatureUncertainty,
    maxT = LandMaxTemperature,
```

```
    maxTU = LandMaxTemperatureUncertainty,
    minT = LandMinTemperature,
    minTU = LandMinTemperatureUncertainty
)

df_global <- df_global %>%
    dplyr::select(-LandAndOceanAverageTemperature, -LandAndOceanAverageTemperatureUncertainty)
```

Skup podataka `df_emission` je problematičan jer se vremenska komponenta izražava u kolonama. To rešavamo tako što transponujemo podatke da bi vremenska komponenta bila vertikalna.

```
library(janitor)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
df_yearly_temps <- df_global %>% group_by(year) %>%
    summarise(temperature = mean(avgT))
df_emission <- as.data.frame(t(as.matrix(df_emission)))
df_emission <- df_emission %>% row_to_names(1)
df_emission_world <- as.data.frame(as.numeric(as.character(df_emission$World)))
colnames(df_emission_world) <- c("world emission")
yearly_emission_and_temp <- cbind(df_yearly_temps, head(df_emission_world, -1))
yearly_emission_and_temp$year <- as.numeric(yearly_emission_and_temp$year)
yearly_emission_and_temp <- yearly_emission_and_temp %>% na.omit()
xts_emission <- as.xts(ts(df_emission, start=1751, frequency = 1, deltat = 1))
```

```
library(stringr)
str2dec <- function(str){
    last_char <- str[nchar(str)-1]
    if(last_char %in% c('N','W'))
    {
        return (as.numeric(str_sub(str, end=nchar(str)-1)))
    }
    else
    {
        return (-1*as.numeric(str_sub(str, end=nchar(str)-1)))
    }
}
df_city$Lng <- df_city$Lng  %>% str2dec()
```

```
## Warning in if (last_char %in% c("N", "W")) {: the condition has length > 1 and
## only the first element will be used
```

```
df_city$Lat <- df_city$Lat  %>% str2dec()
```

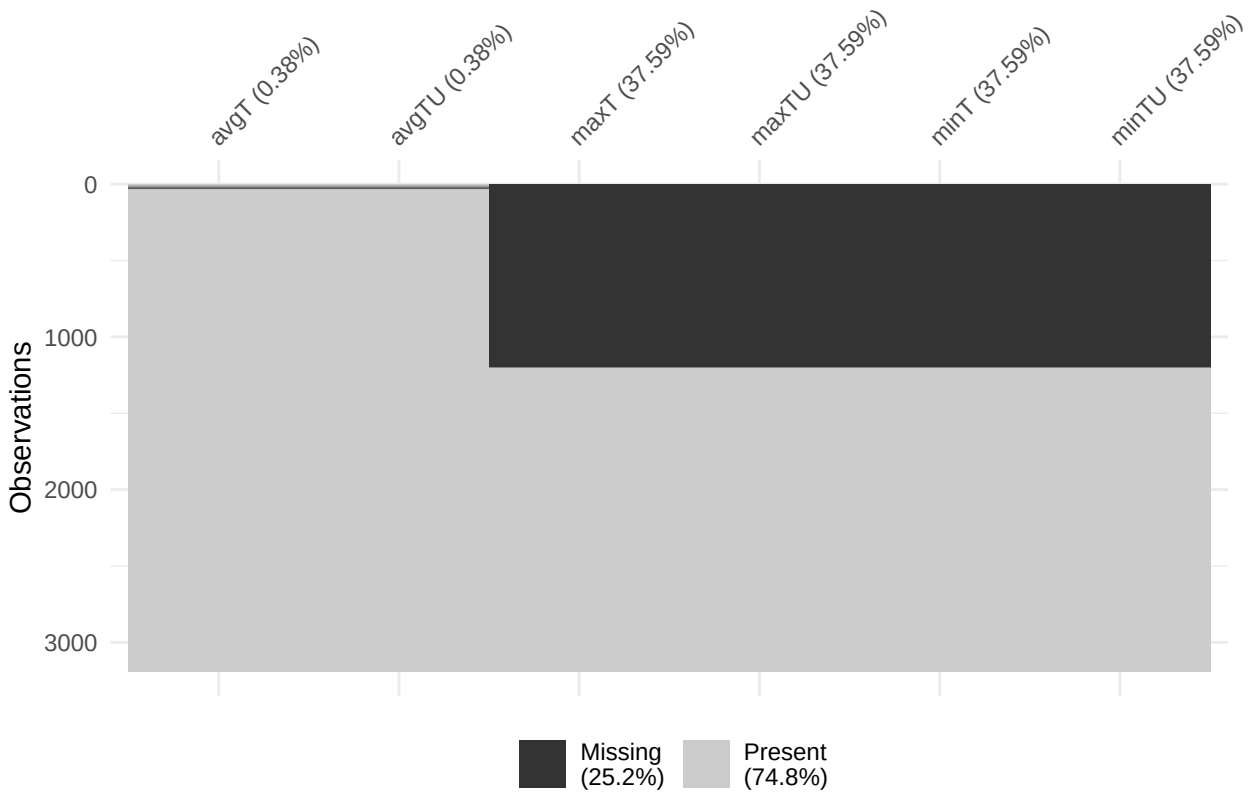
```
## Warning in if (last_char %in% c("N", "W")) {: the condition has length > 1 and
## only the first element will be used
```

```r
head(df_city[c("Lat","Lng")])
```

```
## # A tibble: 6 x 2
##     Lat   Lng
##   <dbl> <dbl>
## 1 -57.0 -10.3
## 2 -57.0 -10.3
## 3 -57.0 -10.3
## 4 -57.0 -10.3
## 5 -57.0 -10.3
## 6 -57.0 -10.3
```

Pronalaženje i rešavanje nepostojećih vrednosti

```r
vis_miss(df_global[2:7])
```



Prilikom pregleda podataka mozemo utvrditi da u `df_global` skupu podataka postoji znacajan broj tj. 25.2% nepostojećih vrednosti. Kako su naše obzervacije hronološki poređane, to se može tumačiti time da se prosečna temperatura meri skoro od početka skupa podataka, dok se minimalna i maksimalna počinju meriti neštko kasnije. Jedo od rešenja jeste vertikalno razdvajanje prosečne i minimalne i maksimalne temperature, no s obzirom da imamo valjan razlog da opravdamo nepostojeće vrednosti, to nećemo uraditi.

Selekcija atributa

```
df_avgT_by_country <- df_country[-3] %>% spread(Country, avgT)
```

Analiza

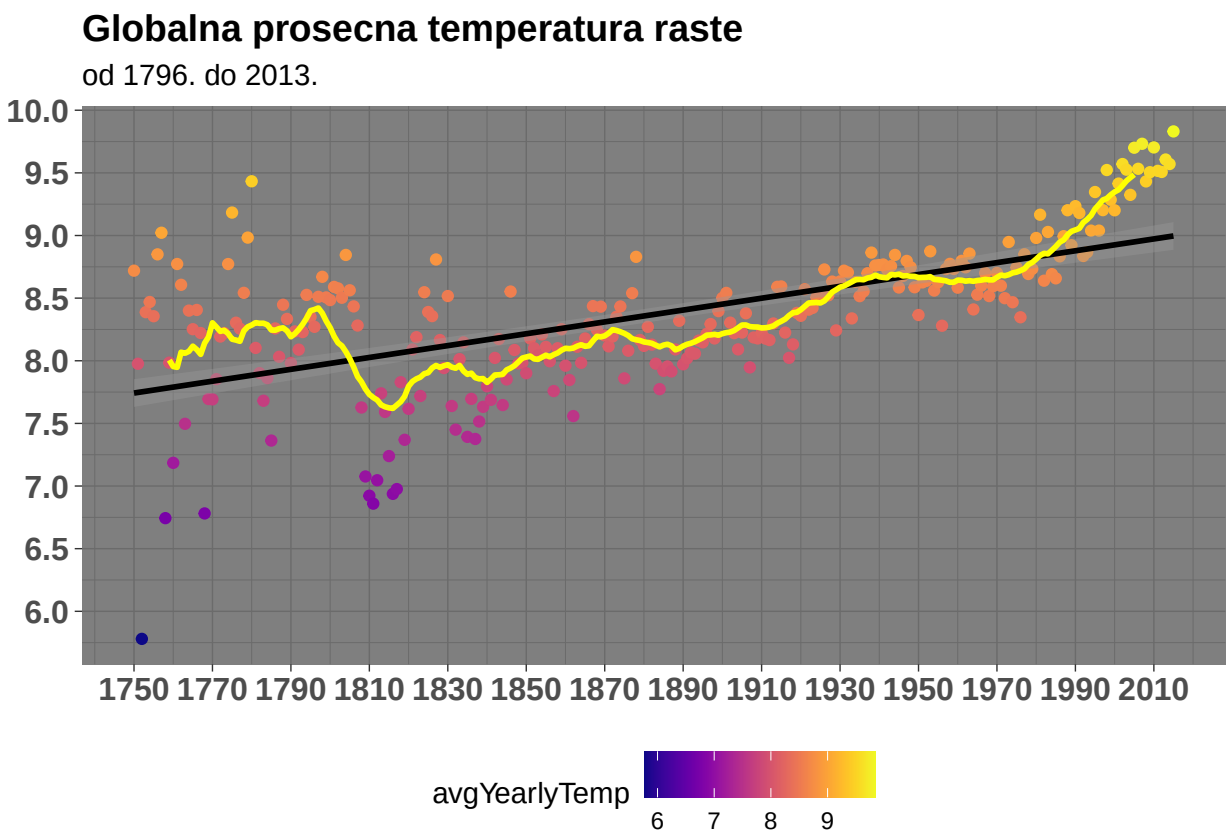
Globalna

Trend

```
data_world <- df_global %>%
    group_by(year) %>%
    summarize(avgYearlyTemp=mean(avgT,na.rm=T))
data_world$year <- as.numeric(data_world$year)

ggplot(data_world, aes(x=year, y=avgYearlyTemp,color=avgYearlyTemp))+
    geom_point()+
    scale_color_viridis(option = "C")+
    geom_smooth(method = "lm", color="black") +
    geom_line(aes(
        y=rollmean(
            avgYearlyTemp, 20,
            na.pad = TRUE)),
        colour="yellow",
        size=1) +
    theme(axis.line = element_line(color = "orange",size=1)) +
    scale_x_continuous(breaks = seq(1750, 2013, by = 20)) +
    scale_y_continuous(breaks = seq(5 , 10, by=0.5)) +
    theme(panel.background=element_blank())+
    theme_dark() +
    theme(legend.position = "bottom",axis.title = element_blank(),
                    axis.text = element_text(size = 12,face="bold"),
        plot.title = element_text(size=14,face = "bold")) +
  ggtitle(sprintf("Globalna prosecna temperatura raste"), subtitle = "od 1796. do 2013.")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



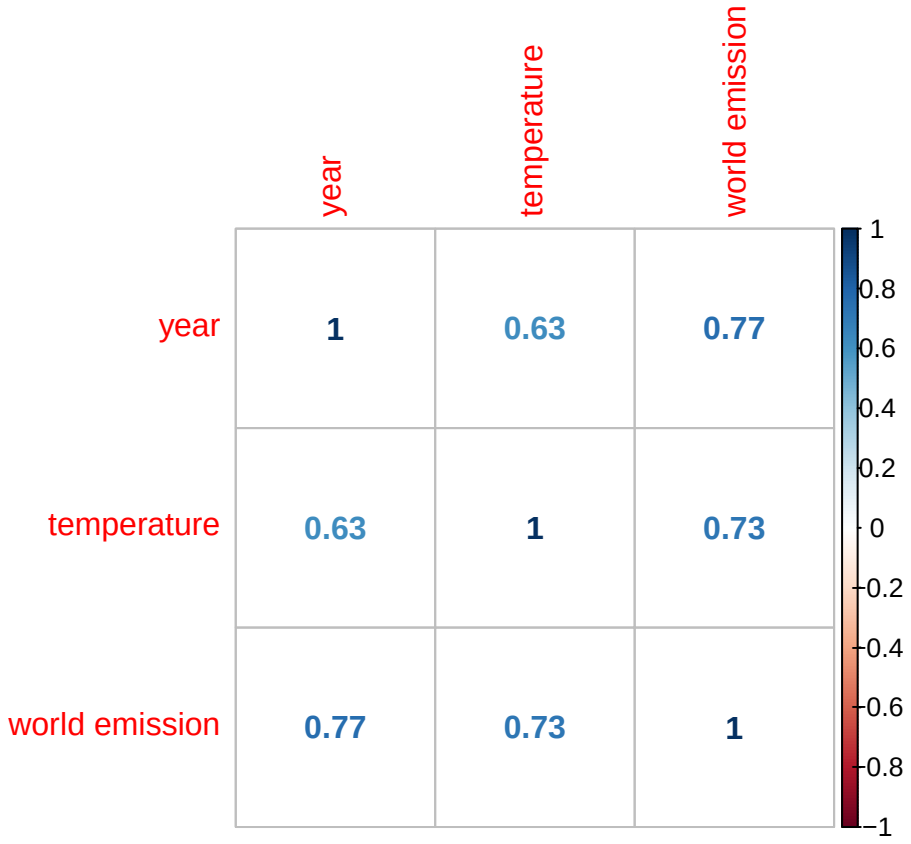
Analiza

- Primetan je porast globalne temperature od oko 1 stepen celzijusa od 1750. godine do danas (crna linija)
- Temperature od 1750. do 1830. godine imaju veliku nesigurnost u merenju
- Od 1975. godine do danas temperatura raste značajnije nego pre (žuta linija)

Uticaj emisije C02 na globalne temperature

Potrebno je ispitati da li emisija ugljen-dioksida utiče na porast globalne temperature. Prvo, treba videti kako izgleda korelaciona matrica.

```
corrplot(cor(yearly_emission_and_temp), method = "number")
```



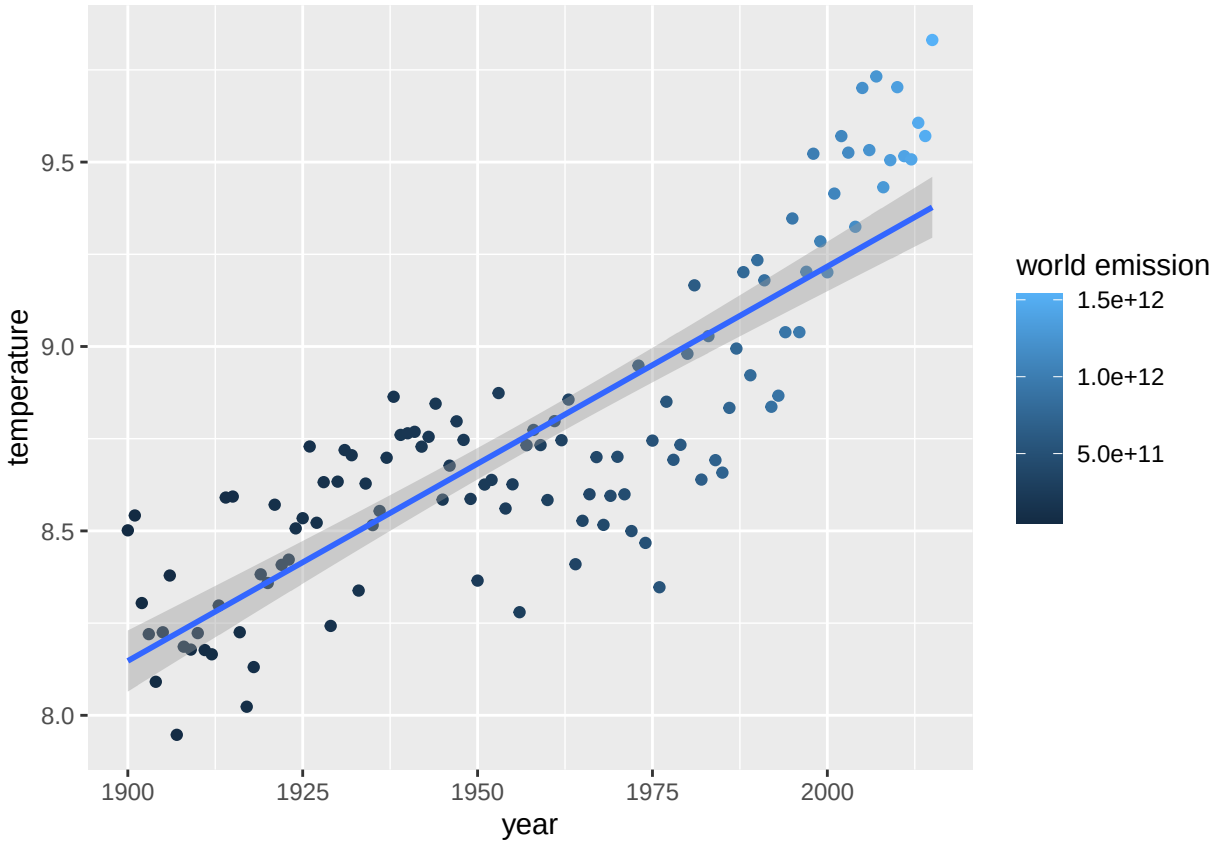
Dalje, možemo da probamo da fitujemo linearan model gde će temperatura zavisiti od emisije ugljen dioksida.

```
df <- yearly_emission_and_temp %>% filter(year >= 1900)
model <- lm(temperature ~ `world emission`, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = temperature ~ `world emission`, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47818 -0.14535  0.00733  0.16225  0.36936
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       8.337e+00  2.694e-02  309.50   <2e-16 ***
## `world emission` 8.998e-13  4.263e-14   21.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1923 on 114 degrees of freedom
## Multiple R-squared:  0.7962, Adjusted R-squared:  0.7944
## F-statistic: 445.4 on 1 and 114 DF,  p-value: < 2.2e-16
```

```
ggplot(data = df, aes(x=`year`, y=temperature)) +
    geom_point(aes(color=`world emission`)) +
    geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



P vrednost modela je višestruko manja od 0.05 iz čega zaključujemo da emisija ugljen dioksida utiče na globalnu temperaturu. S obzirom da je R-squared > 0.7, kažemo da postoji jaka pozitivna veza izmedju emisije CO2 i globalne temperature.

Najpogodjenije države

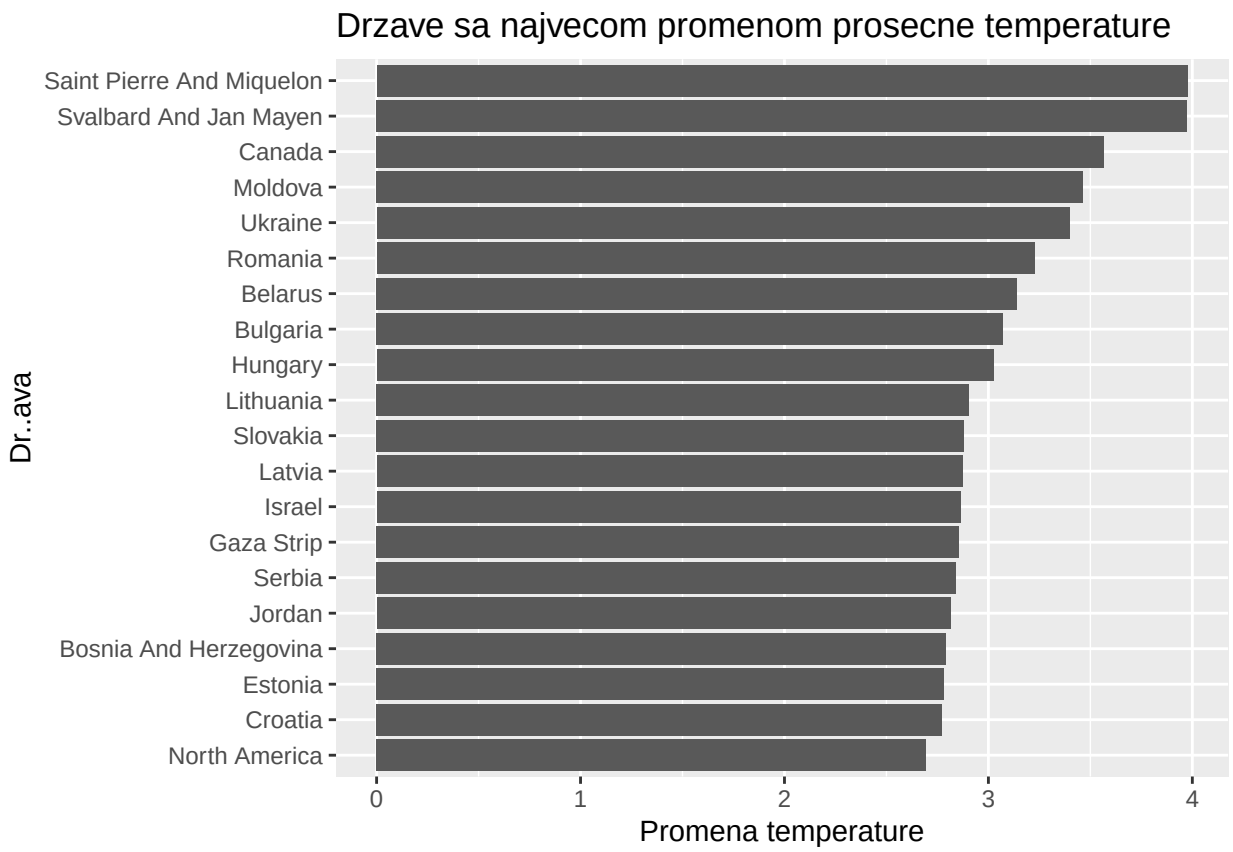
```
dc <- df_country %>%
    filter(year=='1875'|year=='2012')%>%
    group_by(Country,year)%>%
    summarize(temp=mean(avgT))%>%
    spread(year,temp)
```

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

```
dc$change <- dc$`2012`- dc$`1875`
dc <- dc%>%filter(!is.na(change))%>%arrange(desc(change))%>%head(n=20)

dc$Country <- factor(dc$Country, levels = dc$Country)

dc %>%
    ggplot() + geom_col(aes(x=reorder(Country,change), change)) +
    ggtitle("Drzave sa najvecom promenom prosecne temperature") +
    coord_flip() + ylab("Promena temperature") + xlab("Država")
```



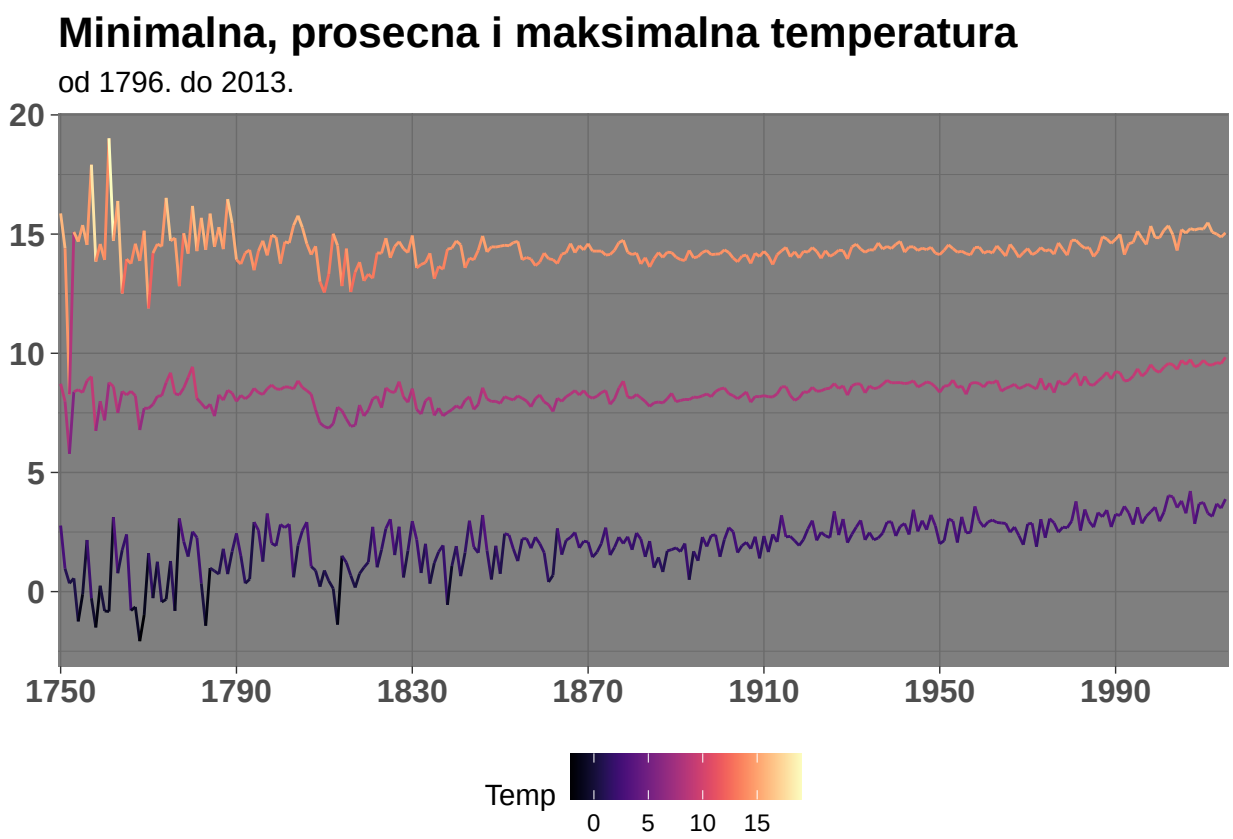
Analiza

- Srbija se nalazi na 15. mestu po promeni prosečne temperature od 1796. godine
- Ta promeni iznosi nešto manje od 3 stepena celzijusa
- Najugroženije su ostrvske zemlje

Minimalna, maksimalna i prosečna temperatura

```
data_globe <- df_global%>%
    group_by(year)%>%
    summarize(max=max(avgT,na.rm=T),min=min(avgT,na.rm=T),Avg_Temp=mean(avgT,na.rm=T)) %>%
    gather(level,Temp, 2:4)

ggplot(data_globe, aes(x=year, y=Temp,colour=Temp,group=level))+
    geom_line() +
    scale_color_viridis(option="A") +
    scale_x_discrete(breaks = seq(1750, 2013, by = 40))+
    theme(panel.background=element_blank())+
    theme_dark()+
    theme(legend.position = "bottom",axis.title = element_blank(),
            axis.text = element_text(size = 12,face="bold"),
            plot.title = element_text(size=16,face = "bold")) +
    ggtitle("Minimalna, prosecna i maksimalna temperatura",subtitle = "od 1796. do 2013.")
```



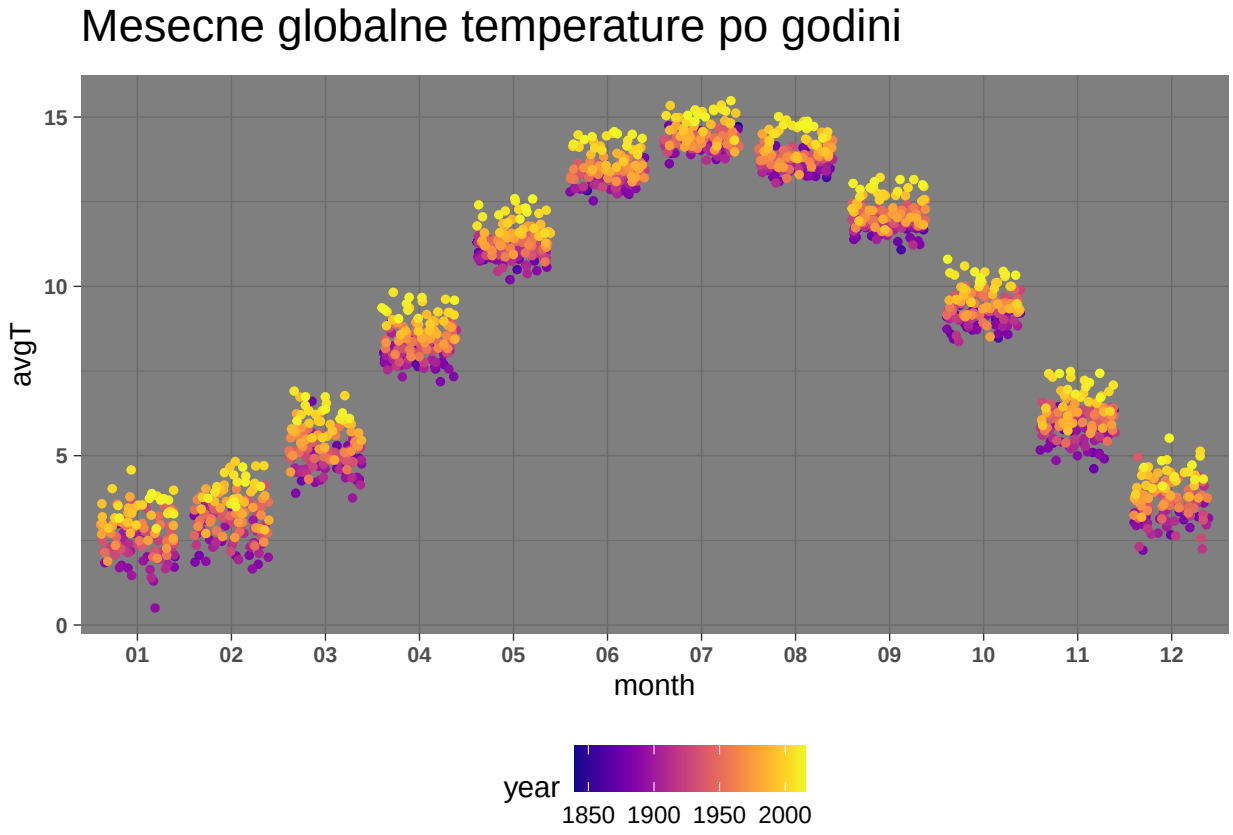
Analiza

- Primećuje se porast prosečne, minimalne i maksimalne temperature po godinama
- Primećuje se i da porast nakon 1950. godine ima veći nagib

Mesečne temperature kroz godine

```
glb_mty <- df_global %>%
    filter(!is.na(avgT)) %>%
    group_by(month) %>%
    filter(avgTU < .5)

ggplot(glb_mty,aes(month,avgT,color=as.numeric(year))) +
    geom_jitter(size=1) +
    scale_color_viridis(option="C")+
    theme(axis.line = element_line(color = "orange",size=.75))+
    theme_dark()+
    scale_x_discrete()+labs(color="year") +
    theme(
        legend.position = "bottom",
        axis.text = element_text(size = 8,face="bold"),
        plot.title = element_text(size=17,face = "bold")) +
    ggtitle(expression("Mesecne globalne temperature po godini"))
```



Analiza

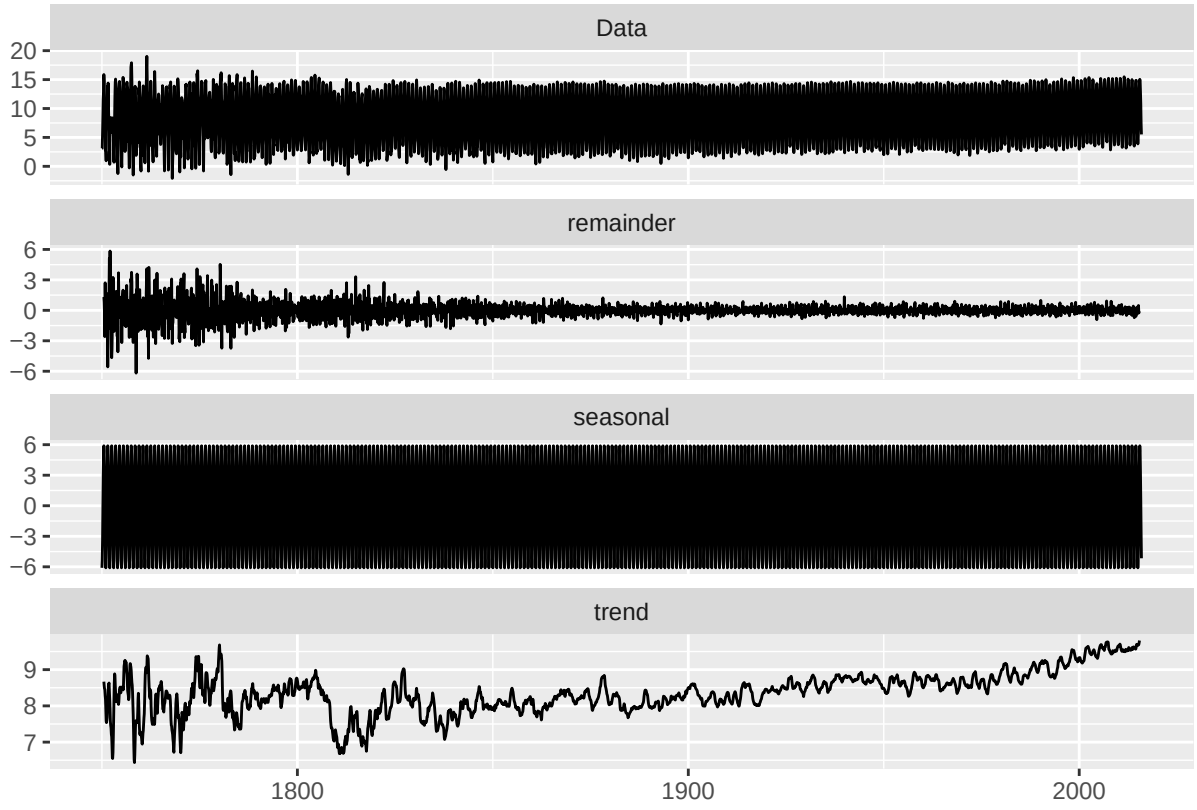
- Sa grafika se može uočiti da se temperature kroz godine za svaki mesec povećavaju
- Takođe se primećuje da je to povećanje značajnije u zimskim mesecima

Dekompozicija vremenske serije

```
ts_global <- ts(na.mean(df_global, option = "mean"), start=c(1750, 1), frequency = 12)

ts_global <- ts_global[,2]
decomp_global_avg <- decompose(ts_global[])

autoplot(decomp_global_avg)
```



Analiza

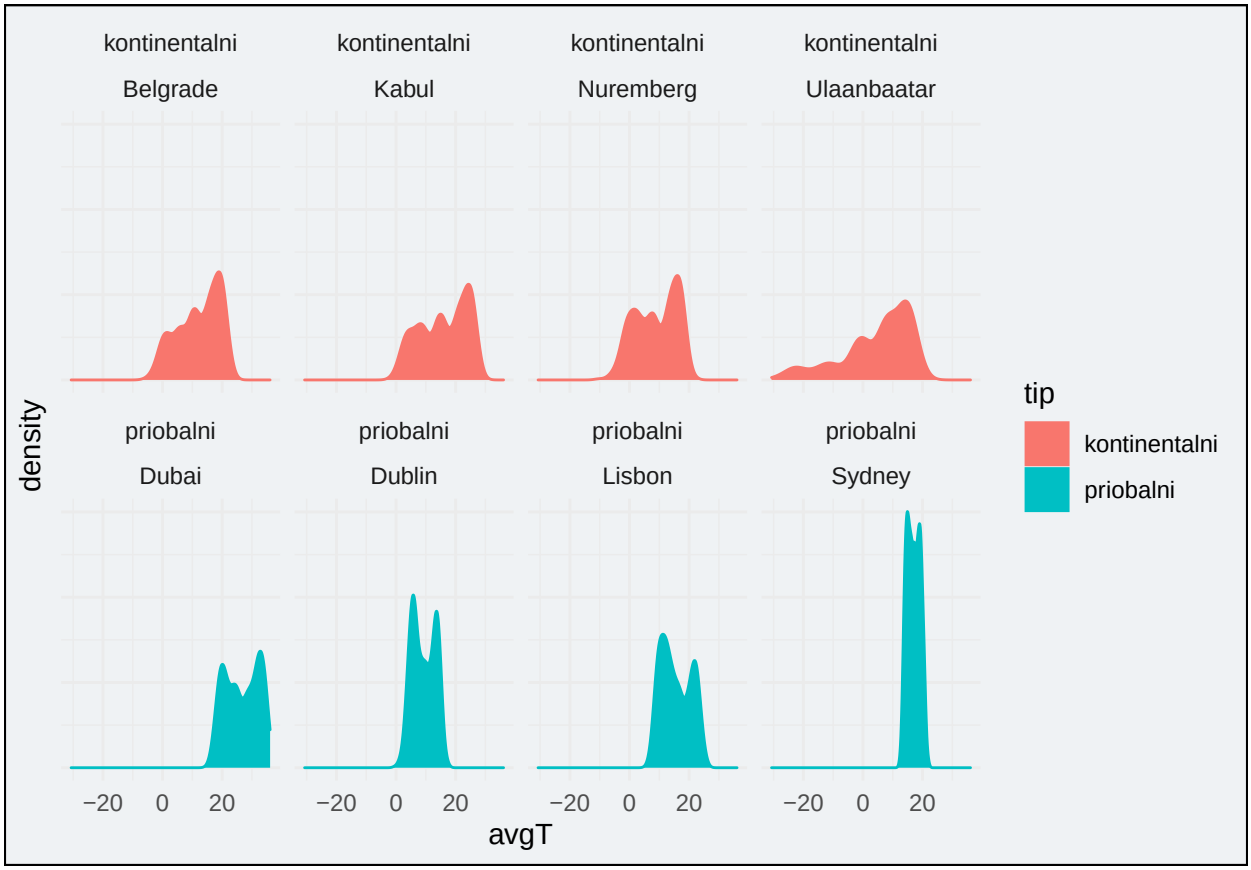
- Globalni trend je pozitivan, zaključujemo da temperature rastu
- Amplituda sezonske komponente iznosi 12 stepena celzijusa

Gradovi

```
coastal_cities = c("Sydney", "Dublin", "Dubai", "Lisbon")
continental_cities = c("Ulaanbaatar", "Nuremberg", "Kabul", "Belgrade")

df_city %>%
    filter(City %in% c(coastal_cities, continental_cities)) %>%
    filter(avgTU <.5) %>%
    mutate(tip=as.factor(ifelse(City %in% coastal_cities, "priobalni", "kontinentalni"))) %>%
    #mutate(time=as.factor(ifelse(dt < as.Date("2000-01-01"), "Pre 2000.", "Posle 2000."))) %>%
```

```
    ggplot(aes(avgT,label=paste(City,""), fill=tip, color=tip, group=tip))+ geom_density() +
    facet_wrap(tip~City, nrow = 2, dir = "h")+
    theme_minimal() +
    theme(
        axis.text.y = element_blank(),axis.ticks.y = element_blank(),
        axis.line.y = element_blank(),strip.background = element_blank(),
        strip.text.y = element_blank(),axis.line.x = element_blank(),
         plot.background = element_rect(fill = "#EFF2F4"),
        plot.title = element_text(size = 14, face = "bold", colour = "gray20", vjust = -1))
```



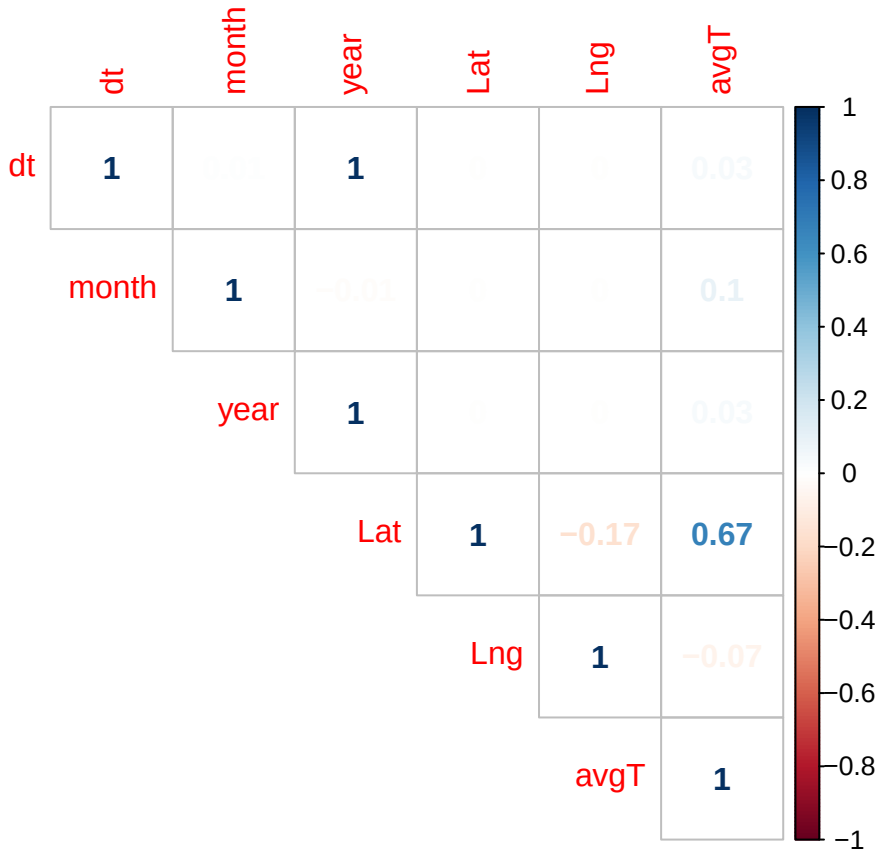
Analiza

- Priobalni gradovi generalno imaju dosta uži raspon temperatura
- Svi kontinentalni gradovi imaju distribuciju koja je nagnuta u desno
- Svi priobalni gradovi imaju distrbuciju koja podseca na slovo `M`

Korelaciona matrica za gradove

```
df <- df_city %>%
    filter(dt >= as.Date("1970-01-01")) %>%
    mutate(dt = as.numeric(as.POSIXct(dt)),
           month = as.numeric(month),
           year = as.numeric(year)) %>% na.omit() %>%
    select(dt, month, year, Lat, Lng, avgT)
```

```
corrplot(cor(df), type = "upper", method = "number")
```



Postoji pozitivna korelacija između geografske širine i izmerene temperature.

Kontinentalna

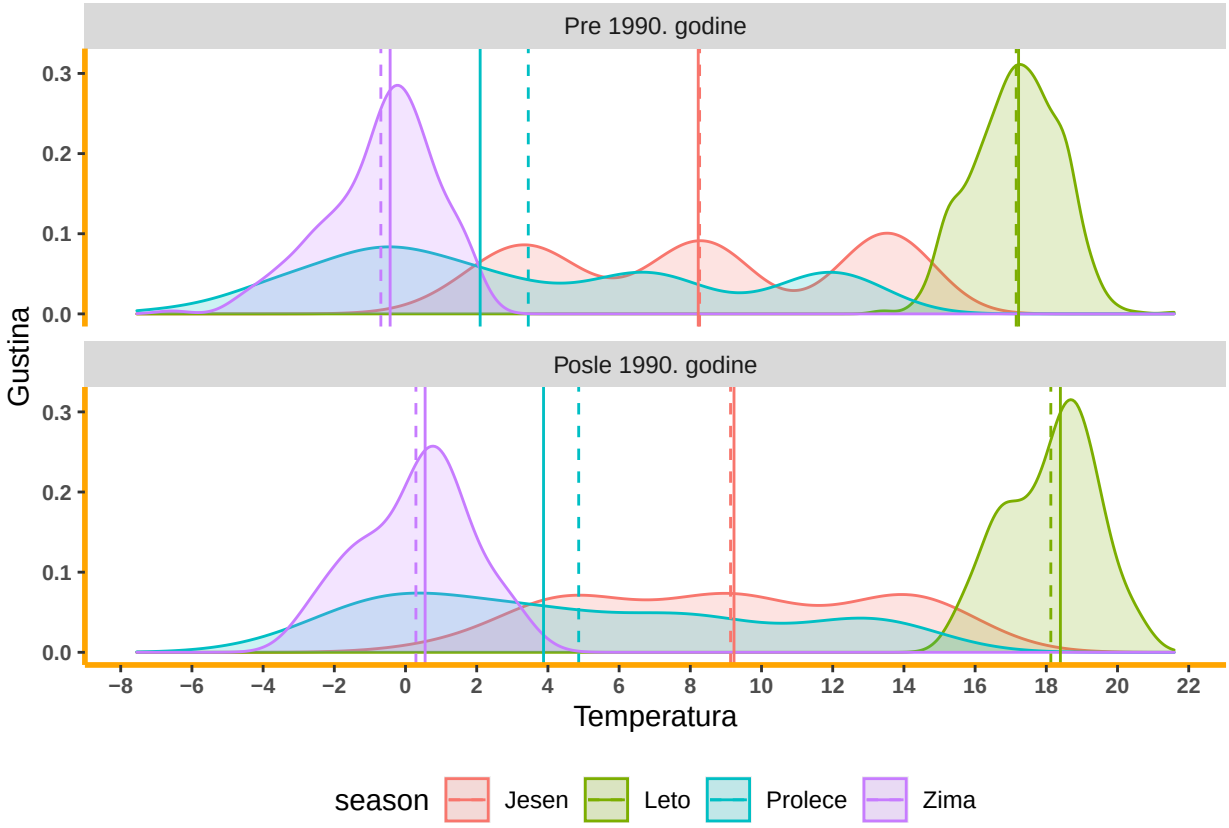
Evropa kroz sezone

```
data_eu <- df_country %>% filter(Country=="Europe") %>% filter(!is.na(avgT))
data_eu$month <- as.integer(data_eu$month)
data_eu <- data_eu %>%
    mutate(season=ifelse(
        month<6,"Prolece",ifelse(month<9,"Leto",ifelse(month<12,"Jesen","Zima")))) %>%
    mutate(before=as.factor(ifelse(dt  >= as.Date("1990-01-01"), TRUE, FALSE)))

levels(data_eu$before) <- c("Pre 1990. godine", "Posle 1990. godine")


ggplot(data_eu,aes(x=avgT))+
    geom_density(aes(group=season,colour=season,fill=season),alpha=0.2)+
    scale_y_continuous(name =  "Gustina")+
    scale_x_continuous(name = "Temperatura", breaks = seq(-10, 24, 2))+
    theme(panel.background=element_blank())+
    theme(axis.line = element_line(color = "orange",size=1))+
```

```
    theme(legend.position = "bottom",
         axis.text = element_text(size = 8,face = "bold"),
        plot.title = element_text(size=12,face = "bold")) +
    stat_central_tendency(aes(color = season), type = "median", linetype = 1)+
    stat_central_tendency(aes(color = season), type = "mean", linetype = 2) +
    facet_wrap(~before, nrow = 2)
```



Analiza

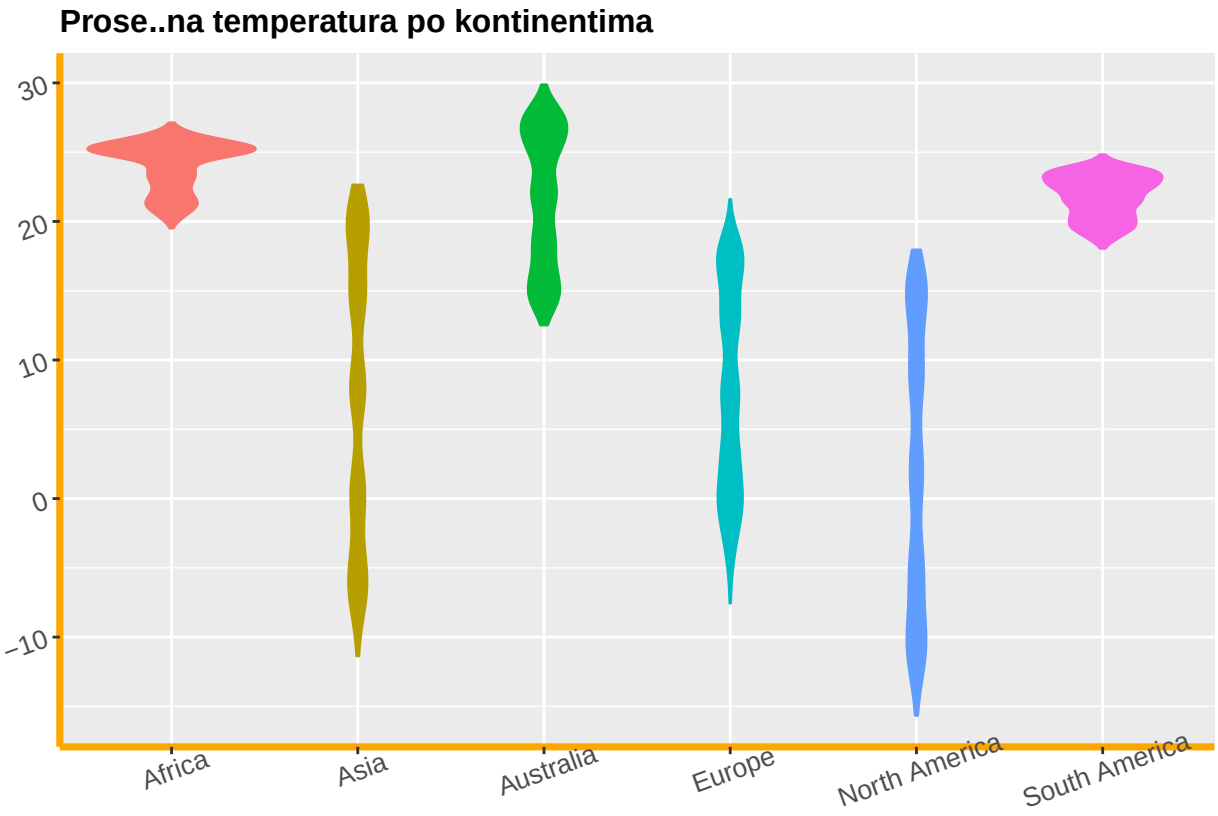
- Primetno je da period posle 1990. godine karakterise porast temperature u svim godišnjim dobima podjednako
- Cela distribucija se pomerila za približno jedan stepen više

Temperature po kontinentima

```
continents <- c("Europe",  "Asia" , "North America", "South America", "Australia", "Africa")
continents_v<-df_country %>% filter(Country %in% continents)%>%filter(!is.na(avgT))

ggplot(continents_v, aes(x=Country,y=avgT,fill=Country, colour=Country))+
    geom_violin()+
    theme(axis.line = element_line(color = "orange",size=1.25))+
    theme(
        legend.position = "none",
        axis.title = element_blank(),
```

```
        axis.text = element_text(size = 10,angle = 20),
        plot.title = element_text(size=12,face = "bold")
    ) +
    ggtitle("Prosečna temperatura po kontinentima")
```



Analiza

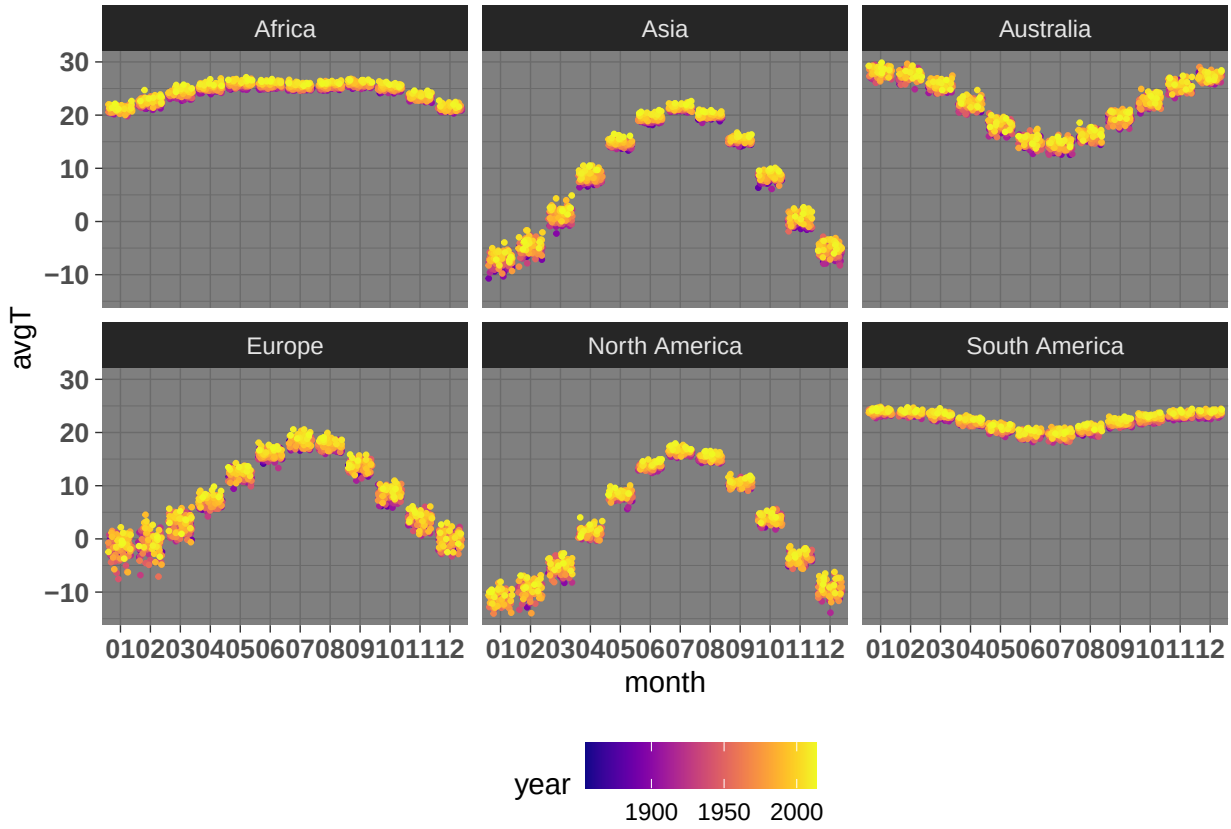
- Afriku i Južnu Ameriku karakteriše mala raširenost distribucije, jer su to kontinenti čija je površina značajnim delom presečena ekvatorom
- Ostali kontinenti imaju znatno širu distribuciju prosečnih temperatura
- Australija dostiže najveće vrednosti zato što je većinom pustinjski kontinent
- Najniže vrednosti dostiže Severna Amerika zbog svoje geografske širine

Mesečne temperature po kontinentima

```
cont_mty <- df_country %>%
    filter(!is.na(avgT)) %>%
    group_by(month) %>%
    filter(avgTU < .5) %>%
    filter(Country %in% continents)

ggplot(cont_mty,aes(month,avgT,color=as.numeric(year))) +
    geom_jitter(size=.5) +
    scale_color_viridis(option="C")+
```

```
    theme(axis.line = element_line(color = "orange",size=.75))+
    theme_dark()+
    scale_x_discrete()+labs(color="year") +
    theme(legend.position = "bottom",
        axis.text = element_text(size = 10,face="bold"),
        plot.title = element_text(size=17,face = "bold")) +
    facet_wrap(~ Country)
```



Analiza

- Evropa, Azija i Severna Amerika imaju raspodelu temperatura po mesecima u obliku zvona. To se tumači time što su sva tri kontinenta na severnoj polulopti.
- Afrika ima raspodelu u obliku slova `M` zato što se ona nalazi svojim delovima i na severnoj i na južnoj polulopti
- Australija i Južna Amerika imaju suprotnu raspodelu od kontinenata koji su na severnoj polulopti

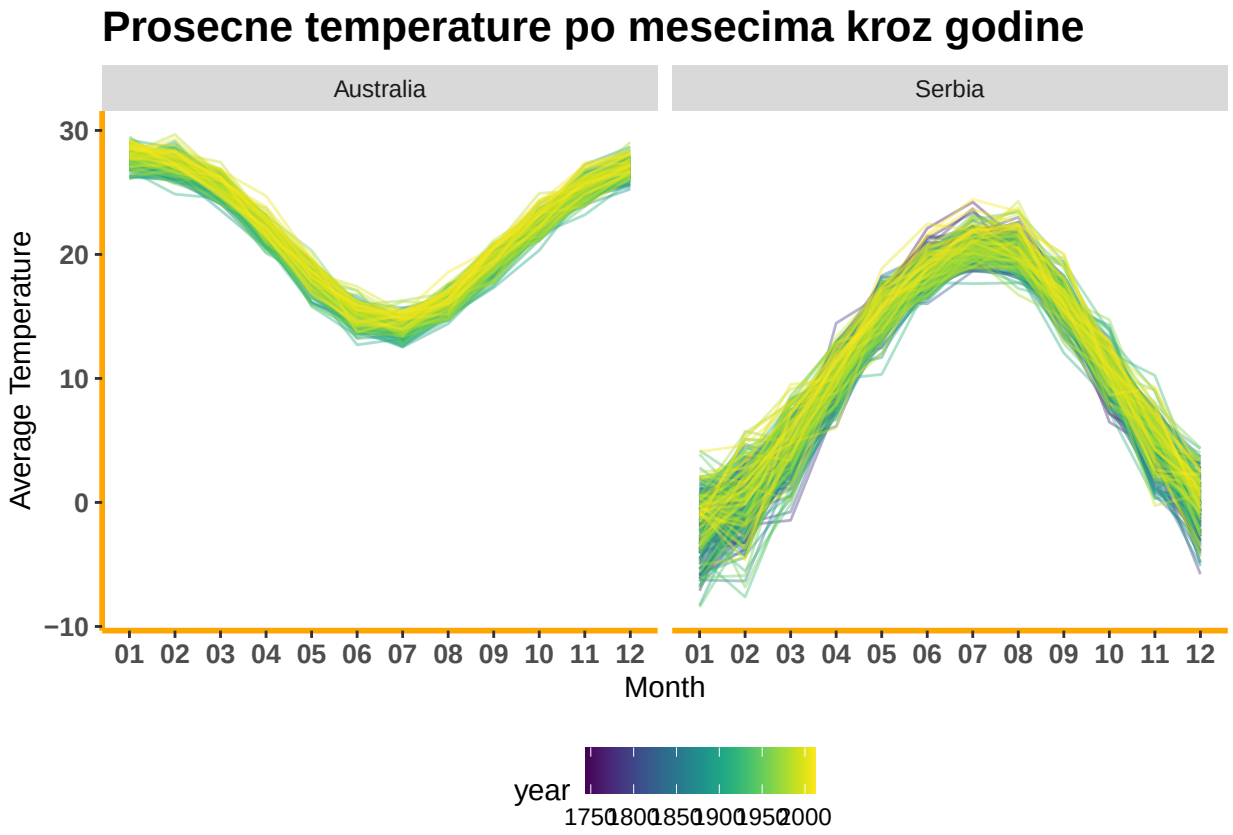
Srbija

Raspodela mesečnih temperatura kroz godine

```
countries <- c("Serbia","Australia")
vals <- df_country %>%
    filter(Country %in% countries) %>%
    group_by(year,Country) %>%
```

```
    mutate(no=length(year)) %>%
    filter(no==12) %>%
    arrange(month)

ggplot(vals,aes(month,avgT, group=year,color =as.numeric(year))) + geom_line(alpha= 0.4) +
  theme(axis.line = element_line(color = "orange",size=1))+
      theme(panel.background=element_blank())+ scale_color_viridis(option="D")+
      scale_x_discrete()+labs(color="year") +
   facet_wrap(~Country)+
    xlab("Month")+ylab("Average Temperature")+
  theme(legend.position = "bottom",
                    axis.text = element_text(size = 10,face="bold"),
        plot.title = element_text(size=16,face = "bold")) +
  ggtitle("Prosecne temperature po mesecima kroz godine")
```



Analiza

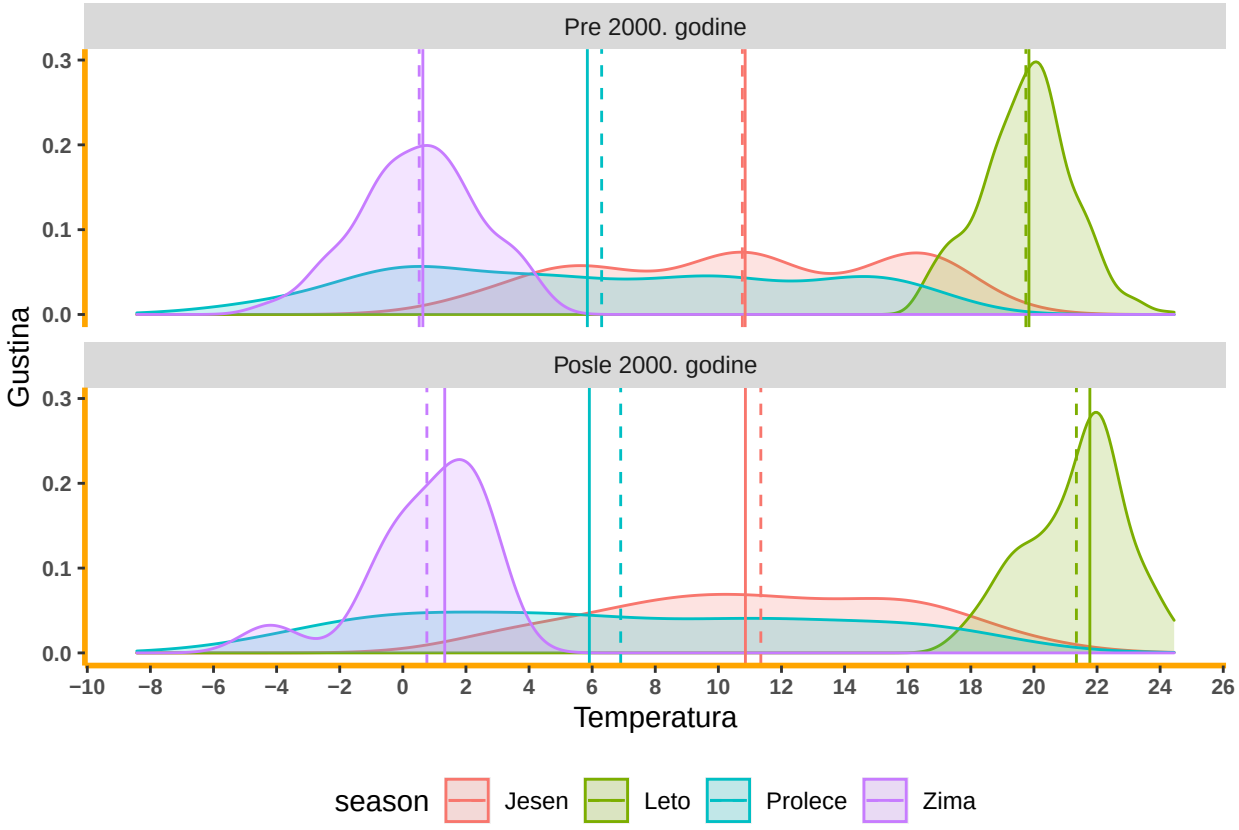
- Srbija u odnosu na Australiju ima veliki raspon temperatura, što se može pripisati kontinentalnoj klimi
- Srbija se nalazi iznad ekvatora, tj. na severnoj polulopti što uzrokuje visoke temperature leti i niske zimi
- Srbija ,kao i Australija, prati trend globalnog porasta temperature

Srbija po sezonama

```
data_srb <- df_country %>% filter(Country=="Serbia") %>% filter(!is.na(avgT)) %>% filter(avgTU < 1)
data_srb$month <- as.integer(data_srb$month)
data_srb <- data_srb %>%
    mutate(
        season=
            ifelse(month<6,"Prolece",
            ifelse(month<9,"Leto",
            ifelse(month<12,"Jesen","Zima")))) %>%
    mutate(before=as.factor(ifelse(dt  >= as.Date("2000-01-01"), TRUE, FALSE)))

levels(data_srb$before) <- c("Pre 2000. godine", "Posle 2000. godine")

ggplot(data_srb,aes(x=avgT))+
    geom_density(aes(group=season,colour=season,fill=season),alpha=0.2)+
    scale_y_continuous(name =  "Gustina")+
    scale_x_continuous(name =  "Temperatura", breaks = seq(-10, 28, 2))+
    theme(panel.background=element_blank())+
    theme(axis.line = element_line(color = "orange",size=1))+
    stat_central_tendency(aes(color = season), type = "median", linetype = 1)+
    stat_central_tendency(aes(color = season), type = "mean", linetype = 2) +
    theme(legend.position = "bottom",
         axis.text = element_text(size = 8,face = "bold"),
        plot.title = element_text(size=12,face = "bold")) +
    facet_wrap(~before, nrow = 2)
```



Analiza

- Leta su postala toplija u poslednjih 20 godina za malo manje od 2 stepena celzijusa kao i zime, za malo manje od 1 stepen celzijusa

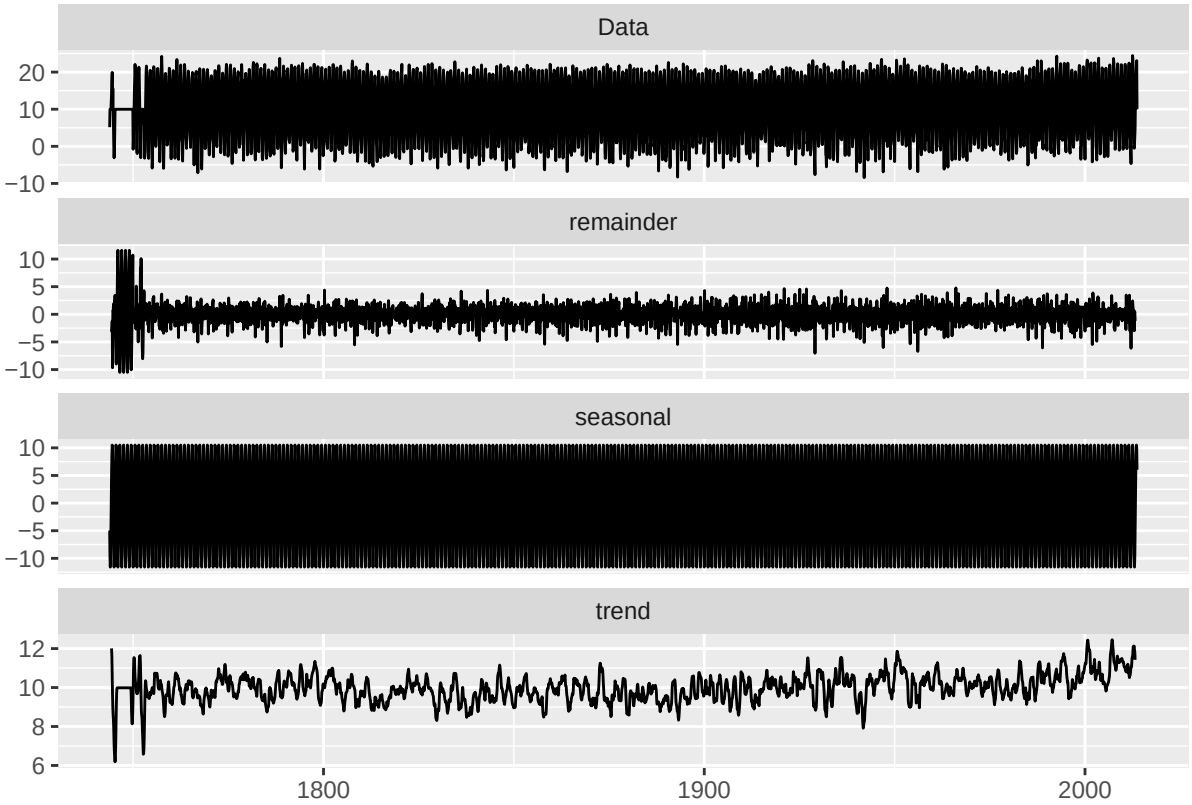
Beograd

Dekompozicija vremeske serije

```
ts_srb <- ts(na.mean(df_country %>% filter(Country == "Serbia"), option = "mean"), start=c(1743, 11), fr

ts_srb <- ts_srb[,2]
decomp_srb<- decompose(ts_srb[])

autoplot(decomp_srb)
```



Analiza

- Srbija prati globalni trend povećanja temperature
- Amplituda vremenske komponente iznosi 20 stepena celzijusa, karakteristično za kontinentalne zemlje

Modelovanje

Formulacija trening i test skupa

```
library(TSstudio)
ts_global_for_modeling <- ts(na.mean(df_global %>% filter(dt >= as.Date("2000-01-01")), option = "mean")
ts_global_for_modeling <- ts_global_for_modeling[,2]
split_ts_global <- ts_split(ts_global_for_modeling, sample.out = 24)
training <- split_ts_global$train
testing <- split_ts_global$test
```

Multivarijabilni generalizovani aditivni model

U ovom koraku je načinjen pokušaj da se srednja mesečna temperatura modeluje preko geografske širine, geografske dužine i meseca u godini. Iako su male šanse da će ovakav model biti adekvatan, nije na odmet pokušati.

```
library(mgcv)
```

```
## Loading required package: nlme
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
## This is mgcv 1.8-31. For overview type 'help("mgcv-package")'.
```

```
df <- df_city %>%
    filter(dt >= as.Date("1970-01-01")) %>%
    mutate(year = as.numeric(year)) %>%
    mutate(month = as.numeric(month)) %>%
    group_by(Lat, Lng, month) %>% summarise(t = mean(avgT)) %>%
    select(t, Lat, Lng, month)
```

```
## `summarise()` has grouped output by 'Lat', 'Lng'. You can override using the `.groups` argument.
```

```
gam_model <- gam(t ~ s(Lat) + s(Lng) + s(month),
                 data = df)
summary(gam_model)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## t ~ s(Lat) + s(Lng) + s(month)
##
## Parametric coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.09470    0.04961   344.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##            edf Ref.df       F p-value
## s(Lat)   8.673  8.961 2736.43  <2e-16 ***
## s(Lng)   8.874  8.995   74.19  <2e-16 ***
## s(month) 7.754  8.605  977.22  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.72   Deviance explained =   72%
## GCV = 32.593  Scale est. = 32.529    n = 13217
```

Ovakav model teško da će adekvatno moći da opiše svu varijansu. Ukupno odstupanje koje je objašnjeno jeste 72% što nije zadovoljavajući rezultat.

SARIMA

```
library(forecast)
```

```
##
## Attaching package: 'forecast'

## The following object is masked from 'package:nlme':
##
##     getResponse

## The following object is masked from 'package:ggpubr':
##
##     gghistogram
```

```
library(tseries)
```

```
##
## Attaching package: 'tseries'

## The following object is masked from 'package:imputeTS':
##
##     na.remove
```

```
adf.test(training, k=12)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  training
## Dickey-Fuller = -2.7525, Lag order = 12, p-value = 0.2622
## alternative hypothesis: stationary
```

```
fcModel <- auto.arima(training, seasonal = T, trace = T)
```

```
##
##  Fitting models using approximations to speed things up...
##
##  ARIMA(2,0,2)(1,1,1)[12] with drift         : 87.01823
##  ARIMA(0,0,0)(0,1,0)[12] with drift         : 210.5186
##  ARIMA(1,0,0)(1,1,0)[12] with drift         : 123.2131
##  ARIMA(0,0,1)(0,1,1)[12] with drift         : 123.3338
##  ARIMA(0,0,0)(0,1,0)[12]                    : 209.2462
##  ARIMA(2,0,2)(0,1,1)[12] with drift         : 110.1829
##  ARIMA(2,0,2)(1,1,0)[12] with drift         : 125.6054
##  ARIMA(2,0,2)(2,1,1)[12] with drift         : 94.68152
##  ARIMA(2,0,2)(1,1,2)[12] with drift         : 89.20962
##  ARIMA(2,0,2)(0,1,0)[12] with drift         : 204.0693
##  ARIMA(2,0,2)(0,1,2)[12] with drift         : 103.3649
##  ARIMA(2,0,2)(2,1,0)[12] with drift         : 103.3585
##  ARIMA(2,0,2)(2,1,2)[12] with drift         : 90.63302
##  ARIMA(1,0,2)(1,1,1)[12] with drift         : 84.32988
##  ARIMA(1,0,2)(0,1,1)[12] with drift         : 107.243
##  ARIMA(1,0,2)(1,1,0)[12] with drift         : 123.5876
##  ARIMA(1,0,2)(2,1,1)[12] with drift         : 95.5999
##  ARIMA(1,0,2)(1,1,2)[12] with drift         : 86.55105
##  ARIMA(1,0,2)(0,1,0)[12] with drift         : Inf
##  ARIMA(1,0,2)(0,1,2)[12] with drift         : 99.36234
##  ARIMA(1,0,2)(2,1,0)[12] with drift         : 102.7843
##  ARIMA(1,0,2)(2,1,2)[12] with drift         : 94.9127
##  ARIMA(0,0,2)(1,1,1)[12] with drift         : 91.46226
##  ARIMA(1,0,1)(1,1,1)[12] with drift         : 82.39043
##  ARIMA(1,0,1)(0,1,1)[12] with drift         : 114.2835
##  ARIMA(1,0,1)(1,1,0)[12] with drift         : 121.452
##  ARIMA(1,0,1)(2,1,1)[12] with drift         : 97.98427
##  ARIMA(1,0,1)(1,1,2)[12] with drift         : 84.58135
##  ARIMA(1,0,1)(0,1,0)[12] with drift         : 203.3763
##  ARIMA(1,0,1)(0,1,2)[12] with drift         : 107.1644
##  ARIMA(1,0,1)(2,1,0)[12] with drift         : 102.595
##  ARIMA(1,0,1)(2,1,2)[12] with drift         : 95.28027
##  ARIMA(0,0,1)(1,1,1)[12] with drift         : 94.2342
##  ARIMA(1,0,0)(1,1,1)[12] with drift         : 83.8527
##  ARIMA(2,0,1)(1,1,1)[12] with drift         : 85.67786
##  ARIMA(0,0,0)(1,1,1)[12] with drift         : 103.7029
##  ARIMA(2,0,0)(1,1,1)[12] with drift         : 83.51136
##  ARIMA(1,0,1)(1,1,1)[12]                    : 81.10814
##  ARIMA(1,0,1)(0,1,1)[12]                    : 114.8773
##  ARIMA(1,0,1)(1,1,0)[12]                    : 119.4759
##  ARIMA(1,0,1)(2,1,1)[12]                    : 96.10984
##  ARIMA(1,0,1)(1,1,2)[12]                    : 83.24198
##  ARIMA(1,0,1)(0,1,0)[12]                    : 201.5775
##  ARIMA(1,0,1)(0,1,2)[12]                    : 106.415
##  ARIMA(1,0,1)(2,1,0)[12]                    : 100.6279
##  ARIMA(1,0,1)(2,1,2)[12]                    : 93.6579
##  ARIMA(0,0,1)(1,1,1)[12]                    : 96.62907
##  ARIMA(1,0,0)(1,1,1)[12]                    : 83.38632
```

```
##  ARIMA(2,0,1)(1,1,1)[12]                    : 84.14451
##  ARIMA(1,0,2)(1,1,1)[12]                    : 83.08829
##  ARIMA(0,0,0)(1,1,1)[12]                    : 106.587
##  ARIMA(0,0,2)(1,1,1)[12]                    : 93.33934
##  ARIMA(2,0,0)(1,1,1)[12]                    : 81.98845
##  ARIMA(2,0,2)(1,1,1)[12]                    : 85.73381
##
##  Now re-fitting the best model(s) without approximations...
##
##  ARIMA(1,0,1)(1,1,1)[12]                    : 79.61675
##
##  Best model: ARIMA(1,0,1)(1,1,1)[12]
```

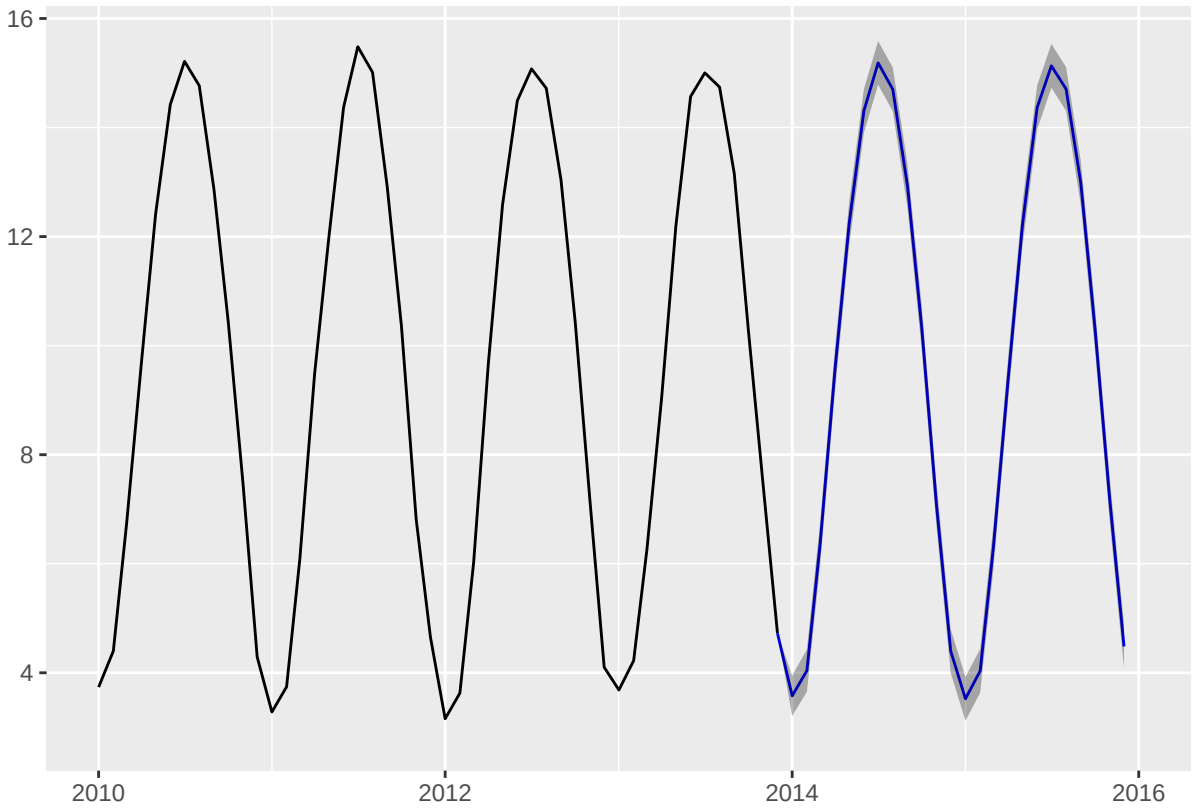
```
summary(fcModel)
```

```
## Series: training
## ARIMA(1,0,1)(1,1,1)[12]
##
## Coefficients:
##          ar1      ma1     sar1     sma1
##       0.6435  -0.3218  -0.2648  -0.8094
## s.e.  0.1438   0.1766   0.0976   0.1072
##
## sigma^2 estimated as 0.08322:  log likelihood=-34.61
## AIC=79.22   AICc=79.62   BIC=94.47
##
## Training set error measures:
##                      ME      RMSE       MAE       MPE     MAPE      MASE
## Training set 0.04988759 0.2743936 0.2105346 0.1967797 3.208522 0.6136658
##                     ACF1
## Training set -0.02230882
```

```
res <- predict(fcModel, n.ahead = 24)
#RMSE na testnom
sqrt(sum((as.data.frame(res$pred) - as.data.frame(testing))^2)) / 24
```

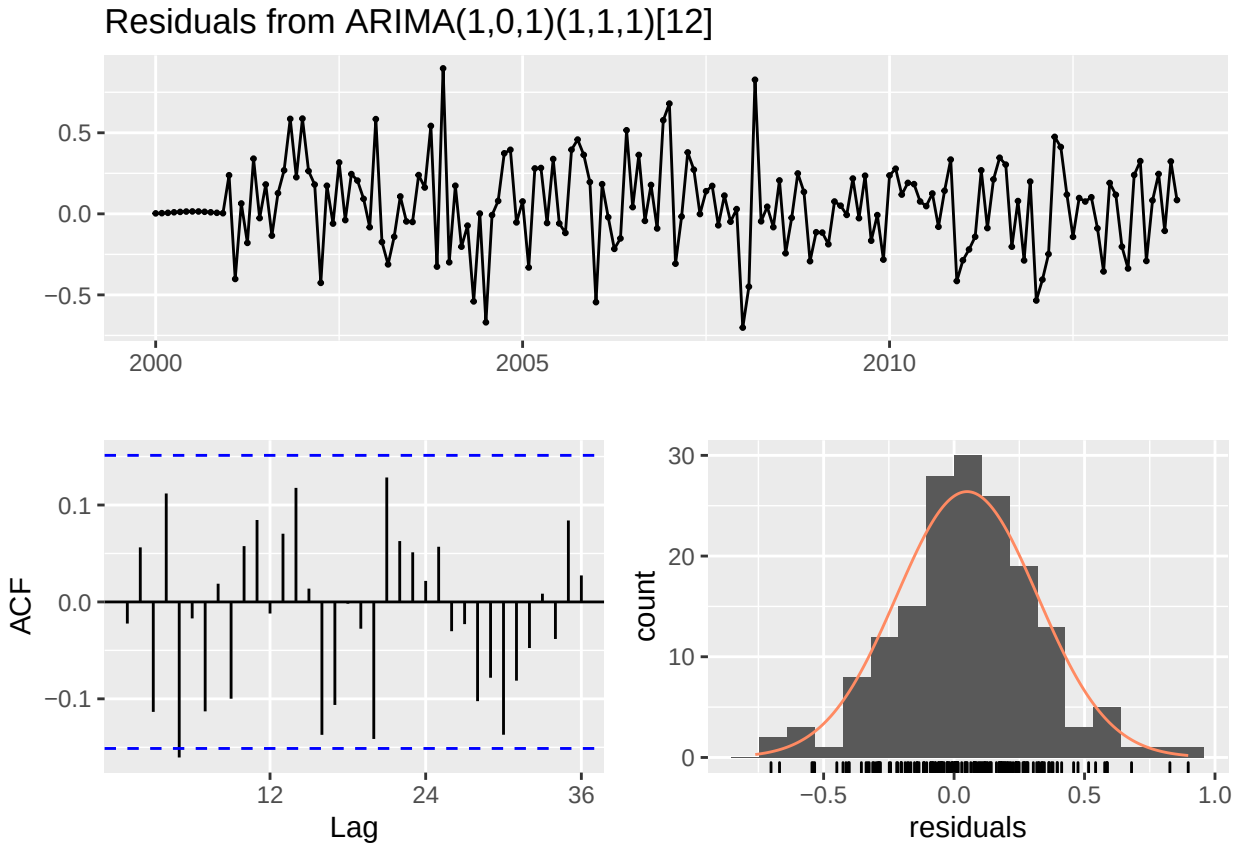
```
## [1] 0.07310078
```

```
autoplot(forecast(fcModel, h = 24)) + xlim(as.Date("2010-01-01"),as.Date("2016-01-01"))
```



Primećujemo da naš model dovoljno dobro predvidja temperaturu u narednih 2 godine.

```
residuals <- checkresiduals(fcModel)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,0,1)(1,1,1)[12]
## Q* = 33.42, df = 20, p-value = 0.03033
##
## Model df: 4.   Total lags used: 24
```

Reziduali nam govore koliko je dobar naš model. Ukoliko su reziduali nisu korelisani i ukoliko im je srednja vrednost bliska 0 onda kazemo da je model dobar. Takodje je poželjno da im je raspodela normalna i da im je varijansa konstantna. Kao što se sa grafika primećuje, ne postoji autokorelacija i raspodela je normalna.

Holt-Wintersovo eksponencijalno glačanje

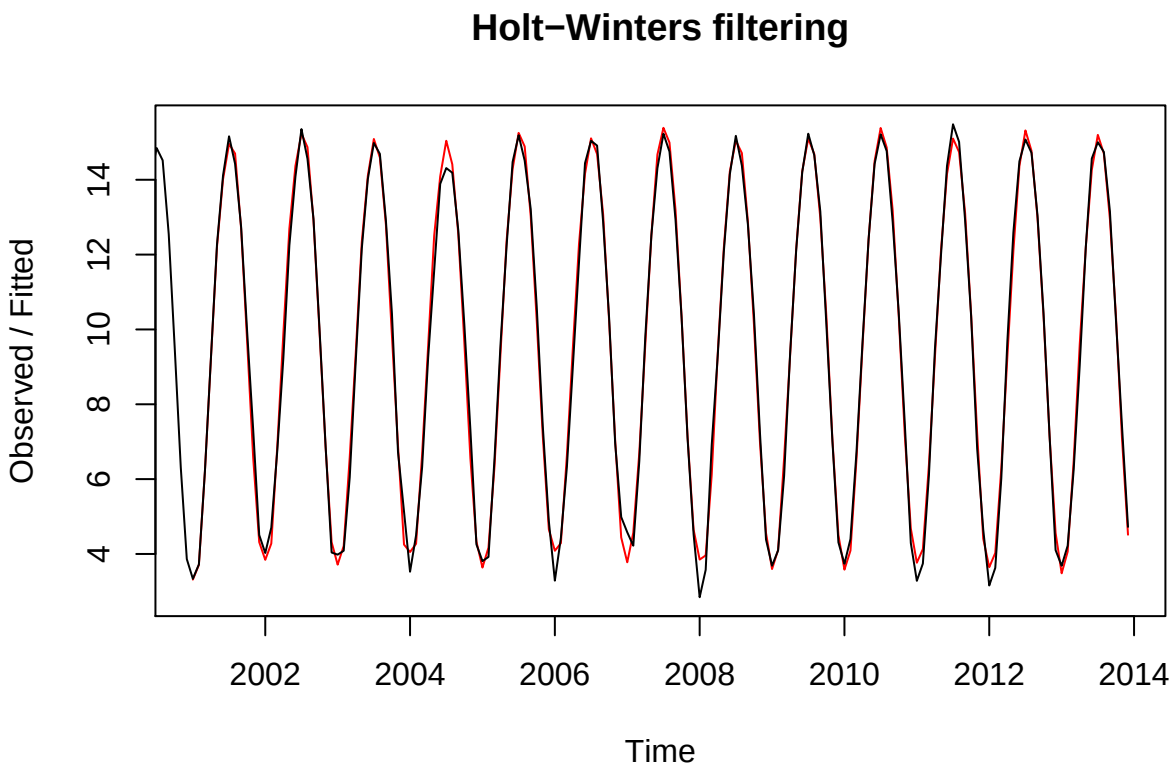
Vremenske serije se mogu takođe modelovati korišćenjem Holt-Wintersovog eksponencijalnog glačanja.

```
temp_timeseries_forcast <- HoltWinters(training)
temp_timeseries_forcast
```

```
## Holt-Winters exponential smoothing with trend and additive seasonal component.
##
## Call:
## HoltWinters(x = training)
##
## Smoothing parameters:
```

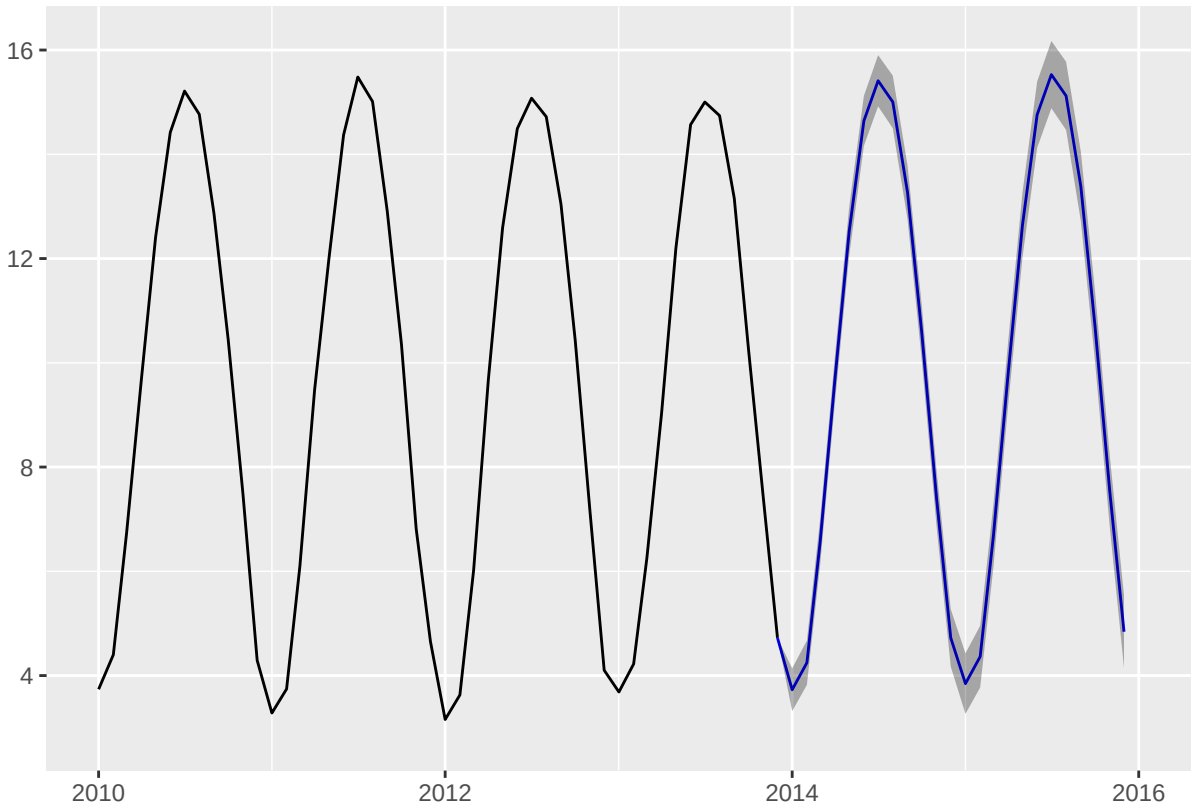
```
##  alpha: 0.265623
##  beta : 0.002688032
##  gamma: 0.2337996
##
## Coefficients:
##            [,1]
## a    9.81210512
## b    0.00979355
## s1  -6.09360840
## s2  -5.58530776
## s3  -3.26806439
## s4  -0.16924326
## s5   2.65541834
## s6   4.77057660
## s7   5.52998989
## s8   5.11237515
## s9   3.36778586
## s10  0.66475500
## s11 -2.50616624
## s12 -5.20623254
```

```
plot(temp_timeseries_forcast)
```




```
forecasted <- forecast(temp_timeseries_forcast, h= 24)
autoplot(forecasted) + xlim(as.Date("2010-01-01"),as.Date("2016-01-01"))
```

```
## Warning: Removed 120 row(s) containing missing values (geom_path).
```



Holt-Winters model daje slične rezultate kao SARIMA.

Zaključak

- Postoji pozitivan trend porasta prosečne globalne temperature
- Postoji jaka veza izmedju emisija ugljen dioksida i porasta globalne temperature
- Postoji pozitivna korelacija izmedju geografske sirine i izmerene temperature
- Kontinentalni i priobalni gradovi imaju drugačiju distribuciju temperature u godini
- Srbija prati globalni trend porasta globalne temperature i nalazi se na 15. mestu po povećanju globalne temperature od 1796. sa porastom nešto manje od 3 stepena

Literatura

- Uvod u programski jezik R, Miloš Ivanović, Tatjana Bošković
- https://medium.com/@kfoofw/seasonal-lags-sarima-model-fa671a858729
- https://r4ds.had.co.nz/exploratory-data-analysis.html
- https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html
- http://environmentalcomputing.net/intro-to-gams/
- https://blog.minitab.com/en/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit