Zaključna naloga

Uvod

Cilj zaključne naloge je, da med reševanjem združite znanje, ki ste ga pridobili tekom predavanj, in ga aplicirate na nek praktičen problem. Naloga simulira vsakodnevni izziv pri delu s podatki, kar pomeni, da bo treba podatke pregledati, prečistiti in strukturirati v bolj primeren format.

Naloga

V mapi *data_raw* je datoteka *weatherAUS.csv*, ki vsebuje meteorološke podatke za mesto Sydney v Avstraliji med 2. julijem 2008 in 25. junijem 2017. Naloga se deli na dva dela: v prvem boste podatke prečistili in shranili v ".xlsx" format, nato pa boste podatke povzeli z vizualizacijami.

1. Del: Čiščenje podatkov

1) Preberite tabelo in odstranite vse stolpce, ki vsebujejo več kot 10% manjkajočih vrednosti (NA).

```
##  [1] "Date"         "Location"     "MinTemp"      "MaxTemp"      "Rainfall"
##  [6] "Evaporation"  "Sunshine"     "WindSpeed9am" "WindSpeed3pm" "Humidity9am"
## [11] "Humidity3pm"  "Pressure9am"  "Pressure3pm"  "Temp9am"      "Temp3pm"
## [16] "RainToday"    "RainTomorrow"
```

2) Ohranite le podatke, ki so bili pridobljeni med vključno "2009-01-01" in "2016-12-31".

```
##           Date Location MinTemp MaxTemp Rainfall Evaporation Sunshine
## 184 2009-01-01   Sydney    18.4    34.7        0         9.8     12.9
## 185 2009-01-02   Sydney    18.8    22.7        0        11.0      5.9
## 186 2009-01-03   Sydney    17.0    23.0        0         9.0      0.5
## 187 2009-01-04   Sydney    18.7    24.6        0         5.4     11.3
## 188 2009-01-05   Sydney    19.5    27.9        0        10.0     12.2
## 189 2009-01-06   Sydney    20.2    28.3        0        10.0     11.8
##     WindSpeed9am WindSpeed3pm Humidity9am Humidity3pm Pressure9am Pressure3pm
## 184           11           17          73          22      1005.5      1000.7
## 185           11           33          62          57      1012.8      1013.8
## 186           13           15          58          49      1021.6      1019.4
## 187           13           31          59          58      1018.6      1015.2
## 188            6           13          66          60      1013.2      1008.9
## 189           11           20          77          64      1011.6      1008.0
##     Temp9am Temp3pm RainToday RainTomorrow
## 184    22.0    34.3        No           No
## 185    20.1    20.8        No           No
## 186    18.8    21.4        No           No
## 187    23.0    24.1        No           No
```

```
## 188    23.4    26.2        No           No
## 189    23.1    25.8        No           No
```

3) Napišite funkcijo, ki za poljuben stolpec zamenja manjkajoče vrednosti s povprečjem prejšnjih dveh dni. Manjkajoče vrednosti v prvih dveh vrsticah naj funkcija zamenja s povprečjem celotnega stolpca. Če stolpec ne vsebuje numeričnih vrednosti naj funkcija vrne napako. Funkcijo aplicirajte na vse numerične in celoštevilske stolpce v tabeli.

```
testni_vector <- c(1, NA, 2, 3, 4, 5, NA)
zapolni_manjkajoce(testni_vector)
```

```
## [1] 1.0 3.0 2.0 3.0 4.0 5.0 4.5
```

4)

a. Napišite funkcijo, ki poljuben vektor **x** transformira s formulo $y = sin(\frac{2\pi}{53}x)$. Nato iz stolpca "Date" izračunajte stolpec, ki bo prikazoval zaporedno število tedna v letu, nad njim ovrednotite prejšnjo funkcijo in ga dodajte v tabelo pod imenom "Weeks".

```
head(data$Weeks, 9)
```

```
## [1] 0.1182732 0.1182732 0.1182732 0.1182732 0.1182732 0.1182732 0.1182732
## [8] 0.2348860 0.2348860
```

b. Postopek ponovite z meseci, toda sedaj z novo formulo $y = cos(\frac{\pi}{6}x)$. Nove podatke shranite v stolpec "Months".

```
head(data$Months, 7)
```

```
## [1] 0.8660254 0.8660254 0.8660254 0.8660254 0.8660254 0.8660254 0.8660254
```

5) Nekatere značilke v tabeli imajo dve meritvi, eno ob devetih zjutraj, drugo ob treh popoldne. Meritvi združite v eno tako, da ohranite le največjo vrednost.

```
##           Date Location MinTemp MaxTemp Rainfall Evaporation Sunshine RainToday
## 184 2009-01-01   Sydney    18.4    34.7        0         9.8     12.9        No
## 185 2009-01-02   Sydney    18.8    22.7        0        11.0      5.9        No
## 186 2009-01-03   Sydney    17.0    23.0        0         9.0      0.5        No
## 187 2009-01-04   Sydney    18.7    24.6        0         5.4     11.3        No
## 188 2009-01-05   Sydney    19.5    27.9        0        10.0     12.2        No
## 189 2009-01-06   Sydney    20.2    28.3        0        10.0     11.8        No
##     RainTomorrow     Weeks    Months WindSpeed Humidity Pressure Temp
## 184           No 0.1182732 0.8660254        17       73   1005.5 34.3
## 185           No 0.1182732 0.8660254        33       62   1013.8 20.8
## 186           No 0.1182732 0.8660254        15       58   1021.6 21.4
## 187           No 0.1182732 0.8660254        31       59   1018.6 24.1
## 188           No 0.1182732 0.8660254        13       66   1013.2 26.2
## 189           No 0.1182732 0.8660254        20       77   1011.6 25.8
```

6) Izpišite kateri dan je bil najbolj vetroven in hkrati tudi deževen.

```
## [1] "2016-06-05"
```

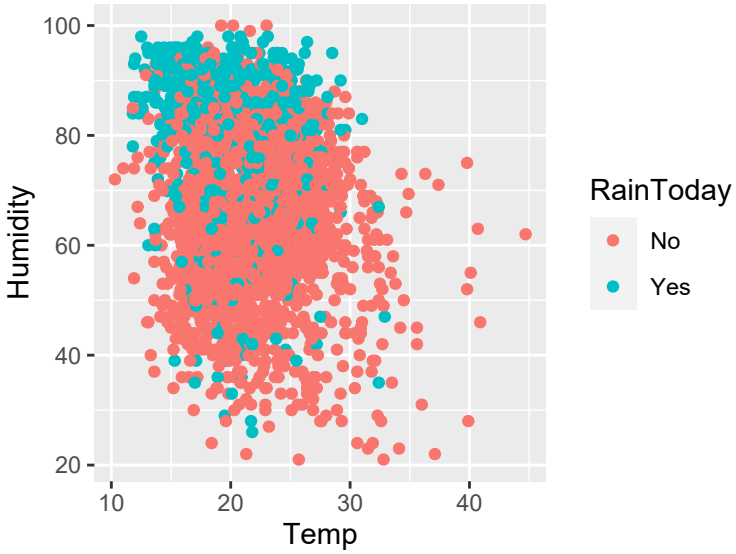
7) Izpišite kolikokrat se je zgodilo, da sta bila dva dneva zapored deževna.

```
## [1] 358
```

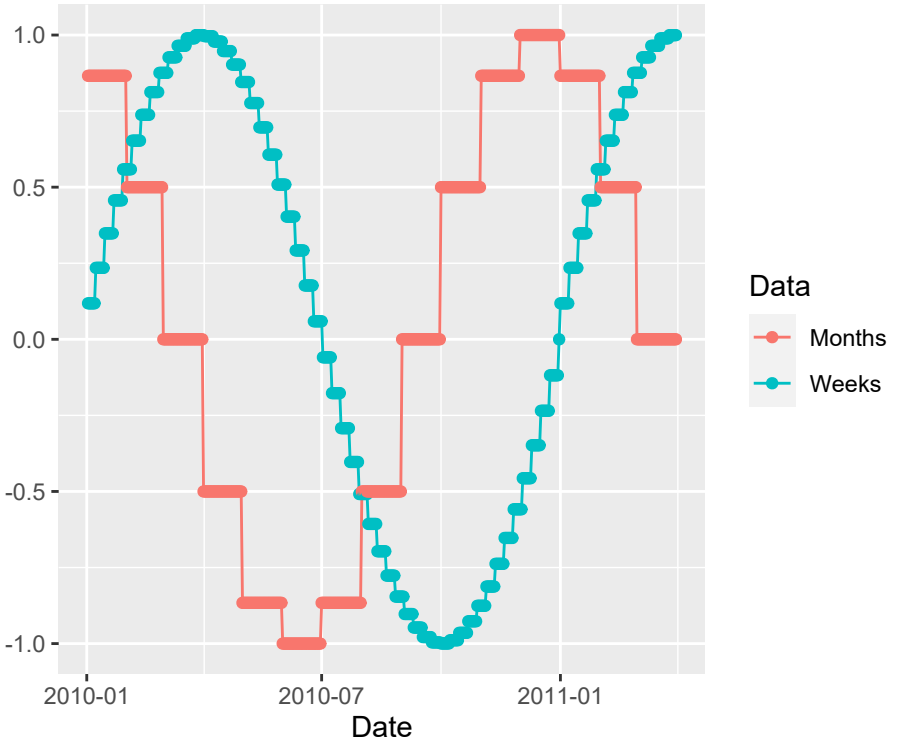
8) Shranite tabelo v .xlsx format z imenom `clean_weatherAUS.xlsx`.

2. Del: Vizualizacije

1) Narišite razsevni diagram, ki prikazuje vlažnost v odvisnosti od temperature. Točke obarvajte različno glede na to, če je tisti dan deževen ali ne.

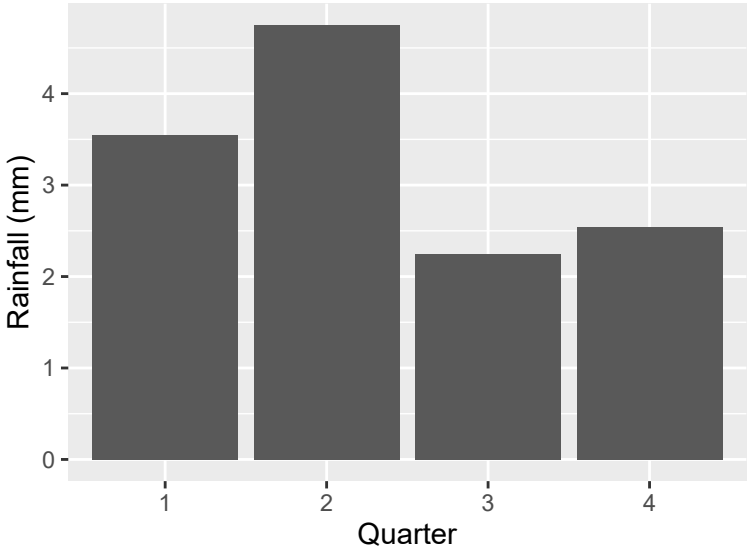


2) S črtnim diagramom prikažite vrednosti stolpcov "Weeks" in "Months" za obdobje med "2010-01-01" in "2011-04-01". Črti naj bosta obarvani glede na stolpec.

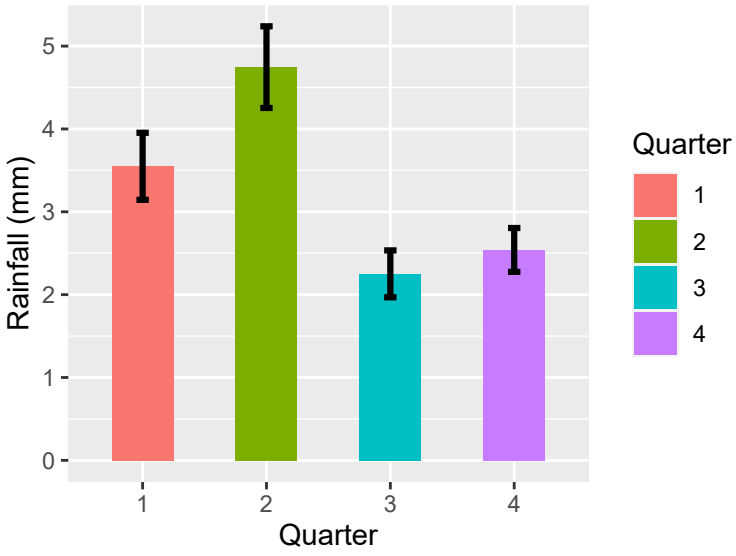


3)

a. S stolpičnim diagramom prikažite povprečne dnevne padavine za letni kvartil.



b. (Težje) Stolpcem dodajte tudi `geom_errorbar`, ki predstavlja standardno napako `error = sd(x)/sqrt(length(x))`.



4. Za stolpce "Pressure", "Humidity" in "Temp" prikažite črtne diagrame s točkami za obdobje med "2012-01-01" in "2013-01-01". Če je bil naslednji dan deževen, točko obarvajte z drugo barvo.