Hausaufgabe Wast3

J.Brändli, T.Haas, F.Kahlbacher

17 Dezember 2018

Inhaltsverzeichnis

	Datum	Zürich Stampfenbachstrasse	Zürich Stampfenbachstrasse_1	Zürich Stampfenbachstrasse_2
1	NA	Zch_Stampfenbachstrasse	Zch_Stampfenbachstrasse	Zch_Stampfenbachstrasse
2	NA	Schwefeldioxid	Kohlenmonoxid	Ozon, höchstes Stundenmittel
3	NA	SO2	CO	O3_max_h1
4	NA	d1	d1	d1
5	NA	µg/m3	mg/m3	µg/m3
6	2012-01-01	1.78	0.31	43.46
7	2012-01-02	1.64	0.24	53.56
8	2012-01-03	1.4	0.24	66.05

Abbildung 1: Header-Zeilen im originalen Datensatz

1 Einleitung

In diesem Bericht werden aktuelle Luftqualitätsmessungen vom Umwelt- und Gesundheitsschutz der Stadt Zürich untersucht. Der Datensatz, welche von der Datenbank OSTLUFT stammt, enthält unteranderem folgende Messgrössen: Schwefeldioxid (SO2), Kohlenmonoxid (CO), Stickstoffdioxid (NO2), Stickstoffmonoxid (NO), Ozonmengen (O3) und meteorolgische Daten, sowie Feinstaubmessungen (PM10). Vor allem die Feinstaubmessungen sind von Relevanz. Der Feinstaub kann tief in die Lunge eindringen und so schweriwiegende Auswirkungen auf die Gesunheit des Menschen haben. Die Messgrössen stammen von den Wetterstationen an der Stampfenbachstrasse, Schimmelstrasse, Heubeeribüelstrasse und der Rosengartenstrasse.

2 Aufgaben

2.1 Aufgabe 1

Der Datensatz wird mit dem R-Befehl *read_csv()* eingelesen. Die eingelesenen Daten werden nun wie gewünscht aufbereitet. Für die Aufbereitung wird unteranderem das R-Packages *dplyr* verwendet, deshalb muss der eingelesene Datensatz in ein Tibble umgewandelt werden.

```
ugz_luftqualitaetsmessung_seit_2012 <- read_csv("ugz_luftqualitaetsmessung_seit-2012.csv")
luftqualitaet <- as_tibble(ugz_luftqualitaetsmessung_seit_2012)
```

Als nächstes werden die Werte und der Header separiert und in neue Variablen abgespeichert. Dieser Vorgang ist nötig, da der orginale Datensatz mehrere Header-Zeilen (siehe Abbildung 1) enthält, die nicht benötigt werden. Zudem vereinfacht dieser Schritt, die Daten in die gewünschte Form aufbereiten zu können.

```
titel <- slice(luftqualitaet, c(2))
titel[1] <- "Datum"
werte <- slice(luftqualitaet, c(-1:-5))
```

Es wird von der Variable *werte* und der Variable *titel* alle Spalten, die zur gleichen Station gehören (Stampfenbachstrasse, Schimmelstrasse, Heubeerbüel und Rosengarten) separiert und zusammengefügt. So erhält man für jede Station den aufbereiteten Datensatz mit ihren Werten und dem gewünschten Header. Am Schluss werden alle Stationen in die Variable *luftqual* zusammengefügt.

```
stampfenbach <- werte %>% select(1,2:14) %>%
  mutate(Station = "Stampfenbachstrasse")
stampfenbach_titel <- titel %>% select(1,2:14)
stampfenbach_titel[length(stampfenbach)] <- "Station"
names(stampfenbach) <- stampfenbach_titel

schimmel <- werte %>% select(1,15:21) %>%
  mutate(Station = "Schimmelstrasse")
schimmel_titel <- titel %>% select(1,15:21)
schimmel_titel[length(schimmel)] <- "Station"
names(schimmel) <- schimmel_titel

heubeer <- werte %>% select(1,22:25) %>%
  mutate(Station = "Heubeeribüel")
heubeer_titel <- titel %>% select(1,22:25)
heubeer_titel[length(heubeer)] <- "Station"
names(heubeer) <- heubeer_titel

rosengarten <- werte %>% select(1,26:30) %>%
  mutate(Station = "Rosengarten")
rosengarten_titel <- titel %>% select(1,26:30)
rosengarten_titel[length(rosengarten)] <- "Station"
names(rosengarten) <- rosengarten_titel


# Tabellen zusammensetzen
luftqual <- bind_rows(stampfenbach, schimmel, heubeer, rosengarten)
```

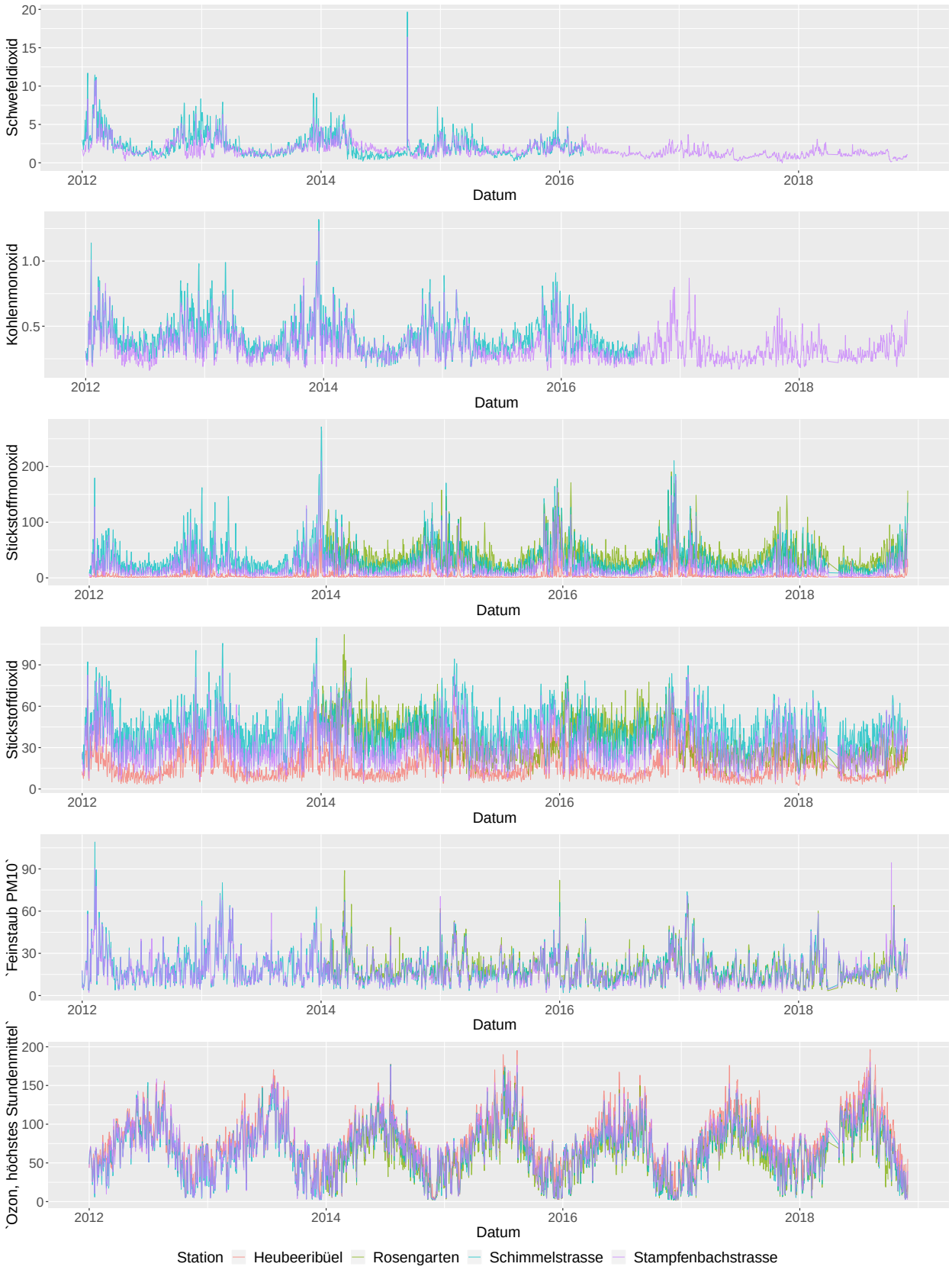
Da das Wetter sich nicht gross zwischen den vier Stationen unterscheidet und es keine näher gelegenen metrologischen Stationen gibt, werden die Wetterdaten der Station Stampfenbachstrasse als Wetterinformation gültig für alle Feinstaubmessstationen.

```
luftqual <- luftqual %>% group_by(Datum) %>% arrange(Datum) %>%
  fill(Lufttemperatur : Regendauer, .direction = "down") %>% arrange(Station)
```

Als letztes werden alle Messwerte in Typ numeric umgewandlet, damit mit diesen gerechnet werden kann. Der aufbereitete Datensatz ist nun in gewünschter Form und kann für die folgenden Aufgaben verwendet werden.

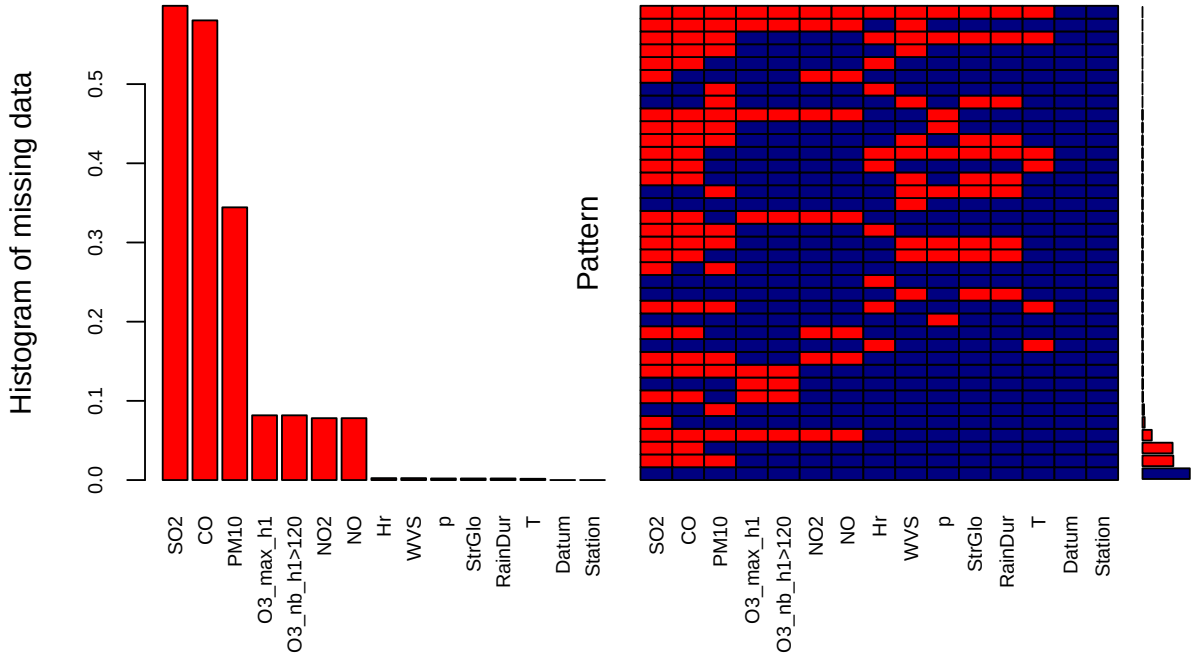
```
luftqual[2:14] <- as_tibble(sapply(luftqual[2:14], as.numeric))
```

2.2 Aufgabe 2


Schwefeldioxid
Kohlenmonoxid
Stickstoffmonoxid
Stickstoffdioxid
`Feinstaub PM10`
`Ozon, höchstes Stundenmittel`
Datum
2012
2014
2016
2018
Station
Heubeeribüel
Rosengarten
Schimmelstrasse
Stampfenbachstrasse

2.3 Aufgabe 3

Im Datensatz gibt es insgesamt 18528 fehlende Werte. In untenstehender Grafik ist links der Anteil der fehlenden Werte pro Variable und rechts der Anteil der fehlenden Werte für Kombinationen von Variablen visualisiert.



Der Anteil der fehlenden Werte pro Variable ist für *SO2* am grössten. Er liegt bei etwa 0.6, dicht gefolgt von *CO* mit 0.58 und *PM10* mit 0.34.

Der Anteil der fehlenden Werte für Kombinationen von Variablen ist schwieriger zu lesen. Die meiste Kombination tritt auf, wenn alle Daten (keine NAs) vorhanden sind. Die zweitmeiste Kombination tritt auf, wenn für SO2, CO2 und PM10 Werte fehlen und alle anderen Variablen keine NAs enthalten.

2.4 Aufgabe 4

In dieser Aufgabe ist die Variable Feinstaub (PM10) von Interesse, jeweils für die Messpunkte Stampfenbachstrasse, Schimmelstrasse und Rosengartenstrasse. Heubeeribüel hat keine Messungen des Feinstaubes und wird darum nicht berücksichtigt.

Der Tagesmittelgrenzwert vom PM10 beträgt 50ug/mˆ3 und darf max. 1x pro Jahr ueberschritten werden.

Daten vorbereiten:

```
luftqual.PM10 <- luftqual %>% ungroup() %>%
  select(Datum, 'Feinstaub PM10', Station) %>%  # nur relevante Kolumnen
  filter(Station != "Heubeeribüel") %>%
  mutate(PM10_uberschritt = `Feinstaub PM10` >= 50 )

head(luftqual.PM10)
```

```
## # A tibble: 6 x 4
##   Datum      `Feinstaub PM10` Station     PM10_uberschritt
##   <date>                <dbl> <chr>       <lgl>
## 1 2012-01-01              NaN Rosengarten NA
## 2 2012-01-02              NaN Rosengarten NA
## 3 2012-01-03              NaN Rosengarten NA
## 4 2012-01-04              NaN Rosengarten NA
## 5 2012-01-05              NaN Rosengarten NA
## 6 2012-01-06              NaN Rosengarten NA
```

Die ersten 6 Werte sind per Zufall gerade NA.

2.4.1 Wie oft wird der Tagesmittel-Grenzwert an welcher Station überschritten?

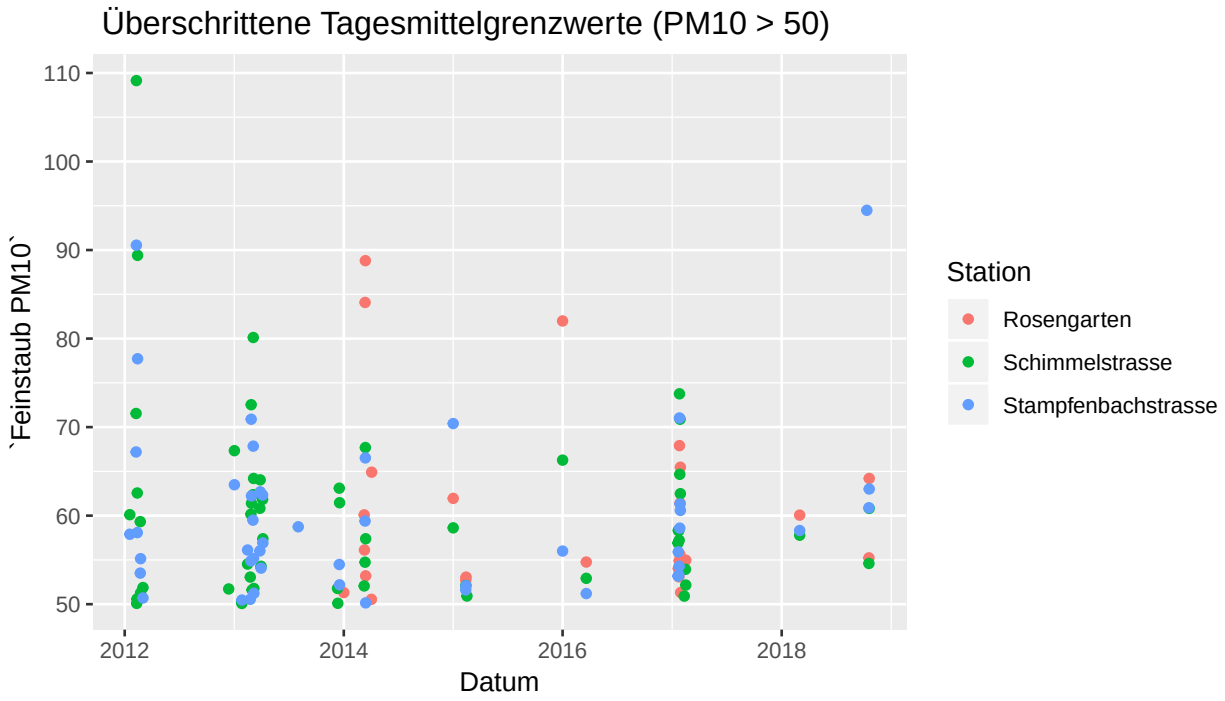
```
luftqual.PM10 %>% group_by(Station) %>% summarize(n = sum(PM10_uberschritt, na.rm = T))
```

```
## # A tibble: 3 x 2
##   Station                n
##   <chr>              <int>
## 1 Rosengarten           25
## 2 Schimmelstrasse       56
## 3 Stampfenbachstrasse   47
```

An der Schimmelstrasse wird der Tagesmittel-Grenzwert am häufigsten überschritten.

2.4.2 Überschrittenen Tagesmittelgrenzwerte im Zeitlichen verlauf für alle Stationen in einer Grafik

```
luftqual.PM10 %>% filter(PM10_uberschritt) %>%
  ggplot(aes(x=Datum, y=`Feinstaub PM10`)) + geom_point(aes(color = Station)) + ggtitle(" Überschritten
```



Der Tagesmittelgrenzwert wird bei allen Stationen oft in den gleichen Wochen / Monaten uebertroffen. Dies ist gut moeglich, da alle Messstationen ähnlichen Einflüssen ausgesetzt sind, auf welche der Feinstaub reagiert. Wie z.B. die Jahreszeit und das Wetter.

Aus der Grafik ist ersichtlich, dass die Feinstaubbelastung im Winter höher ist, als in den anderen Jahreszeiten.

2.4.3 In welchen Jahren und Stationen ist der Anteil der Tage mit Grenzwert uberschreitungen signifikant groesser als zufaellig

Jahresmittelgrenzwert = 20ug/mˆ3

Daten vorbereiten: Nach dem Jahr und Station gruppieren und Werte mit NA entfernen.

```
luftqual.PM10.2 <- luftqual.PM10 %>%
  mutate(Jahr = strtrim(luftqual.PM10$Datum, 4)) %>%
  drop_na() %>%
  group_by(Jahr, Station) %>%
  summarize(n = sum(PM10_uberschritt, na.rm = T))
```

Statistischer Test, ob die Anzahl Überschreitungen signifikant sind. Wir verwenden den Binom.test, da es sich um mehrere Bernoulli Experimente handelt (0 keine Überschreitung, 1 Überschreitung).
h0: eine überschreitungen pro Jahr
h1: mehr als 1 überschreitungen pro Jahr

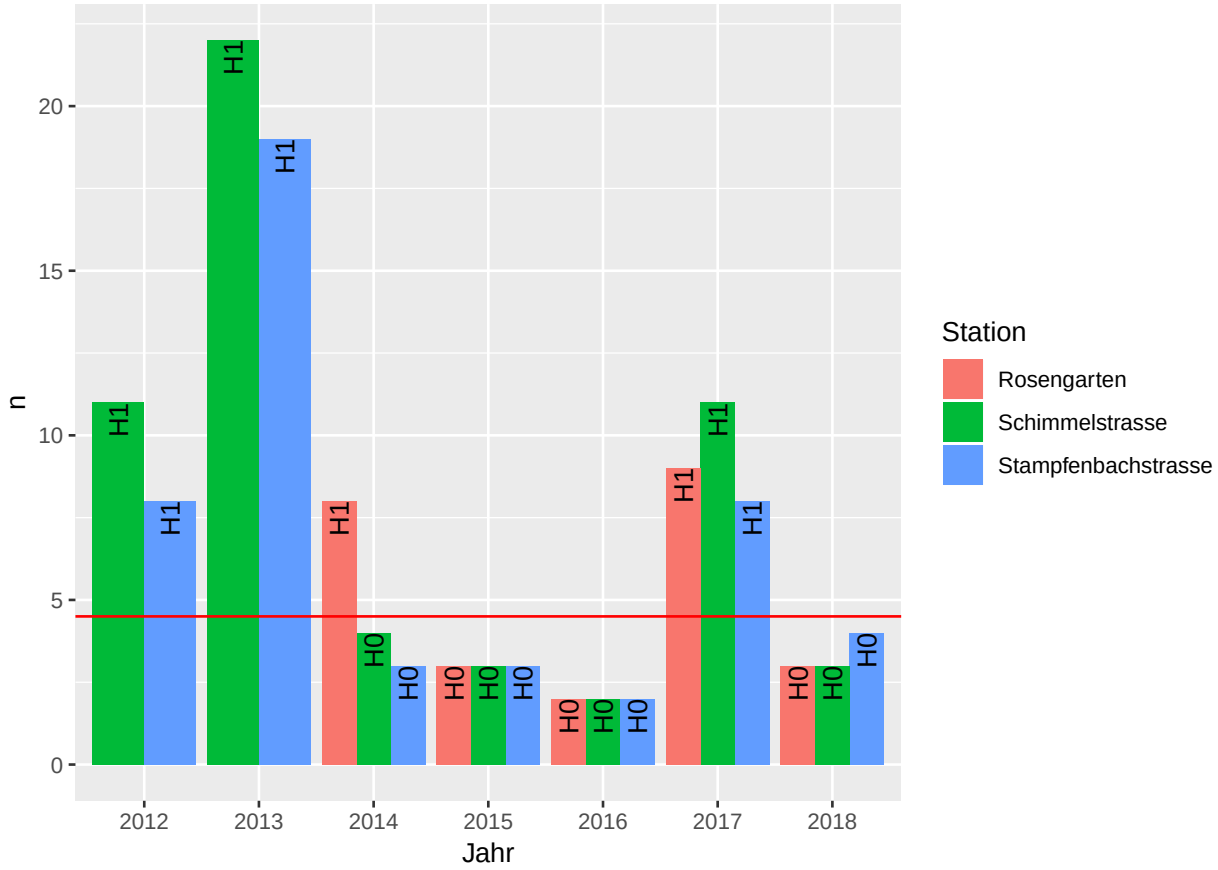
```
fun_A4 <- function(anzahl){
  testen <- binom.test(x = anzahl, n = 365, p = 1/365, alternative = "greater", conf.level = 0.99)
  return(ifelse(testen$p.value<0.01, "H1", "H0"))
}
luftqual.PM10.2$h <- apply(luftqual.PM10.2[,3], 1, fun_A4)
luftqual.PM10.2
```

```
## # A tibble: 19 x 4
## # Groups:   Jahr [?]
##    Jahr  Station                n h
##    <chr> <chr>              <int> <chr>
##  1 2012  Schimmelstrasse        11 H1
##  2 2012  Stampfenbachstrasse     8 H1
##  3 2013  Schimmelstrasse        22 H1
##  4 2013  Stampfenbachstrasse    19 H1
##  5 2014  Rosengarten             8 H1
##  6 2014  Schimmelstrasse         4 H0
##  7 2014  Stampfenbachstrasse     3 H0
##  8 2015  Rosengarten             3 H0
##  9 2015  Schimmelstrasse         3 H0
## 10 2015  Stampfenbachstrasse     3 H0
## 11 2016  Rosengarten             2 H0
## 12 2016  Schimmelstrasse         2 H0
## 13 2016  Stampfenbachstrasse     2 H0
## 14 2017  Rosengarten             9 H1
## 15 2017  Schimmelstrasse        11 H1
## 16 2017  Stampfenbachstrasse     8 H1
## 17 2018  Rosengarten             3 H0
## 18 2018  Schimmelstrasse         3 H0
## 19 2018  Stampfenbachstrasse     4 H0
```

h0 wird ab 5 überschreitungen im Jahr verworfen.

Plot der Überschreitungen

```
ggplot(luftqual.PM10.2, aes(x = Jahr, y = n, fill = Station)) +
  geom_col(aes(), position = "dodge") +
  geom_text(aes(label=h),angle = 90, position = position_dodge(0.9), hjust = 1, size = 4)+
  geom_hline(yintercept = 4.5, col = "red")
```



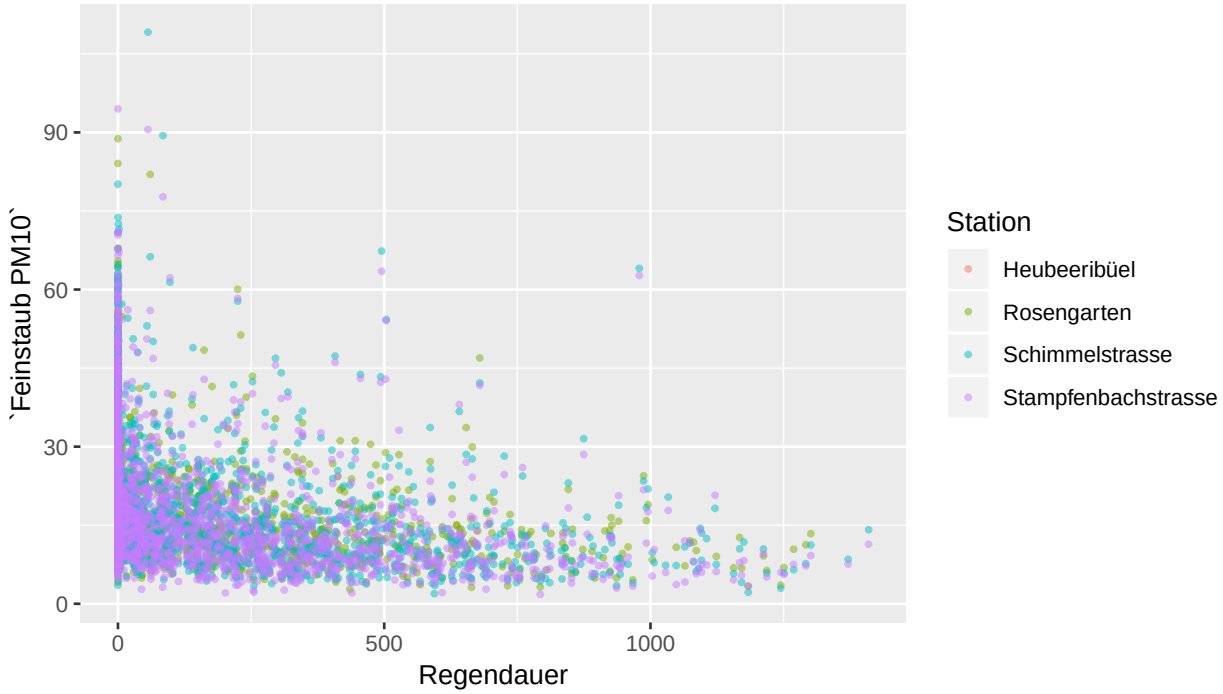
Die Jahren und Stationen bei welchen die Balken über der roten Linie liegen, bei diesen wird die H0 verworfen. Bei allen die darunter liegen wird H0 angenommen.

2.5 Aufgabe 5

Visualisierung des Zusammenhanges zwischen der Regenmenge und der Feinstaubkonzentration pro Station.

```
ordered_PM10 <- select(luftqual, Datum, `Feinstaub PM10`, Station, Regendauer)

# Plot
ggplot(ordered_PM10, aes(x= Regendauer, y= `Feinstaub PM10`)) +
  geom_point(aes(color = Station),  alpha = 0.5, size = 0.8)
```



Auf der Grafik ist folgendes Verhalten zu erkennen: Die Feinstaubkonzentration ist tendenziell höher wenn es nicht Regnet, als wenn es Regnet. Um so länger es Regnet, nimmt dieser Effekt noch zu.

Mit einem Statistischen Test soll überprüft werden, ob ein Zusammenhang besteht. Es wurde der t-Test verwendet, da die Mittelwerte der Feinstaubkonzentration normalverteilt aber die Varianz nicht bekannt sind.

h0: mu_Feinstaub_regen = mu_Feinstaub_keinregen
h1: mu_Feinstaub_regen < mu_Feinstaub_keinregen

```
PM10_test <- ordered_PM10 %>% group_by(Datum) %>%
  summarise(PM10_mean = mean(`Feinstaub PM10`, na.rm = T) ,
            Regendauer = Regendauer[1])

# tTest
t.test(x = PM10_test$PM10_mean[PM10_test$Regendauer != 0],
       y = PM10_test$PM10_mean[PM10_test$Regendauer == 0],
       conf.level = 0.99, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  PM10_test$PM10_mean[PM10_test$Regendauer != 0] and PM10_test$PM10_mean[PM10_test$Regendauer ==
## t = -16.808, df = 2027.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 99 percent confidence interval:
##       -Inf -6.078394
## sample estimates:
## mean of x mean of y
##  16.08592  23.14162
```

p < 0.01 –> h0 verwerfen

Die Nullhypothese wird klar verworfen. Das heisst, die Feinstaubkonzentration ist an Tagen mit Regen signifikant tiefer als an Tagen mit Regen.

2.6 Aufgabe 6

3 Fazit