

Hausaufgabe Wast3

J.Brändli, T.Haas, F.Kahlbacher

17 Dezember 2018

Inhaltsverzeichnis

1	Einleitung	2
2	Aufgaben	2
2.1	Aufgabe 1	2
2.2	Aufgabe 2	4
2.3	Aufgabe 3	5
2.4	Aufgabe 4	6
2.5	Aufgabe 5	8
2.6	Aufgabe 6	9
3	Fazit	9

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 unter anderem folgende Messgrößen: Schwefeldioxid (SO₂), Kohlenmonoxid (CO), Stickstoffdioxid (NO₂), Stickstoffmonoxid (NO), Ozonmengen (O₃) und meteorologische Daten, sowie Feinstaubmessungen (PM₁₀). Vor allem die Feinstaubmessungen sind von Relevanz. Der Feinstaub kann tief in die Lunge eindringen und so schwerwiegende Auswirkungen auf die Gesundheit des Menschen haben. Die Messgrößen stammen von den Wetterstationen an der Stampfenbachstrasse, Schimmelstrasse, Heubeeribühlstrasse 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 unter anderem das R-Packages `dplyr` verwendet, deshalb muss der eingelesene Datensatz in ein Tibble umgewandelt werden.

```
ugzluftqualitaetsmessung_seit_2012 <- read_csv("ugzluftqualitaetsmessung_seit-2012.csv")
luftqualitaet <- as_tibble(ugzluftqualitaetsmessung_seit_2012)
```

Als nächstes werden die Werte und der Header separiert und in neue Variablen abgespeichert. Dieser Vorgang ist nötig, da der originale 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.

	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	SO ₂	CO	O ₃ _max_h1
4	NA	d1	d1	d1
5	NA	µg/m ³	mg/m ³	µg/m ³
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

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

2.1 Aufgabe 1

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

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

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

rosengarten <- wert<e %>% 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.

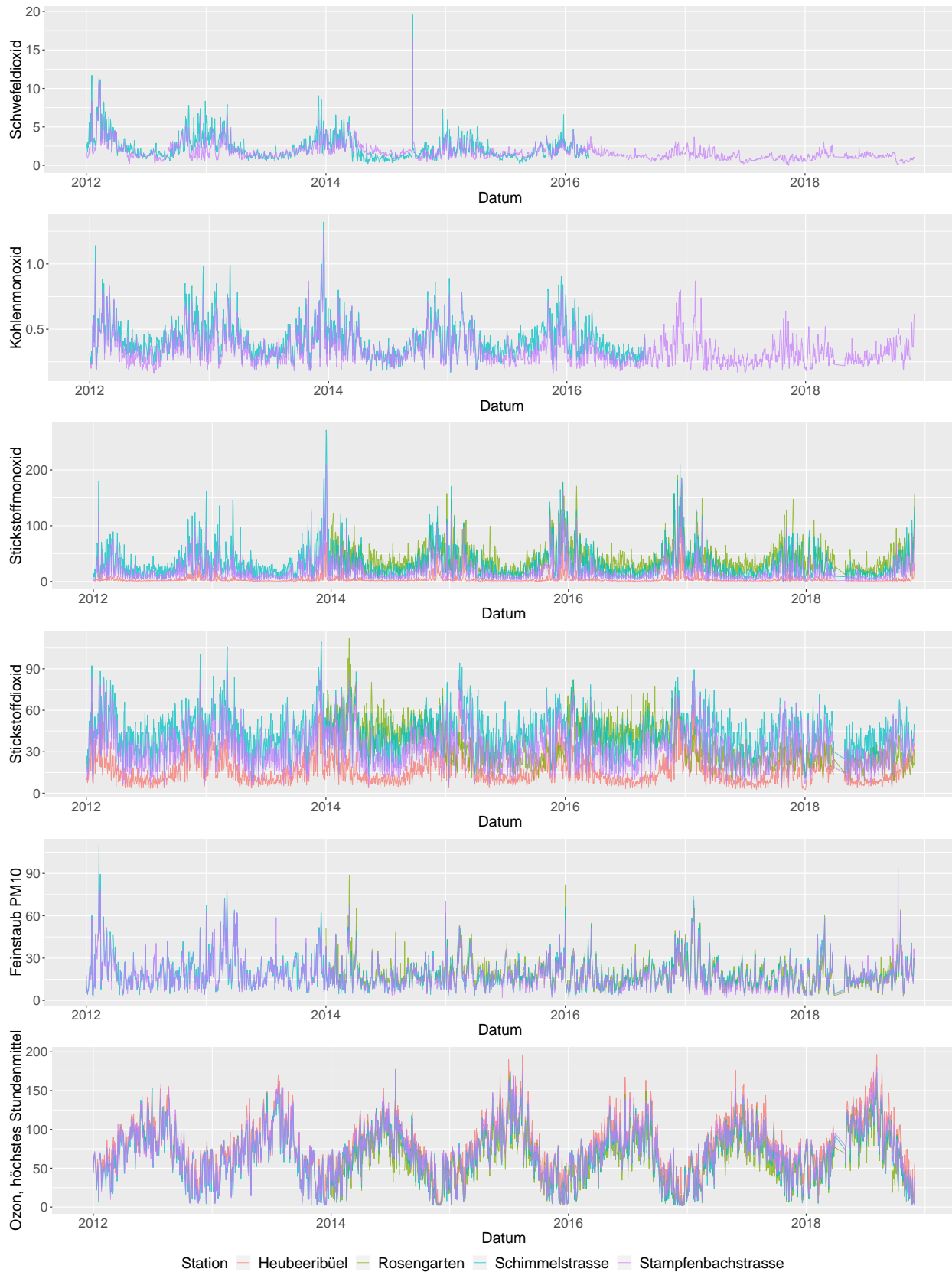
```
luftqual <- luftqual %>% group_by(Datum) %>% arrange(Datum) %>%
  fill(Lufttemperatur : Regendauer, .direction = "down") %>% arrange(Station)
```

Als letztes werden alle Messwerte in Typ numeric umgewandelt, 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

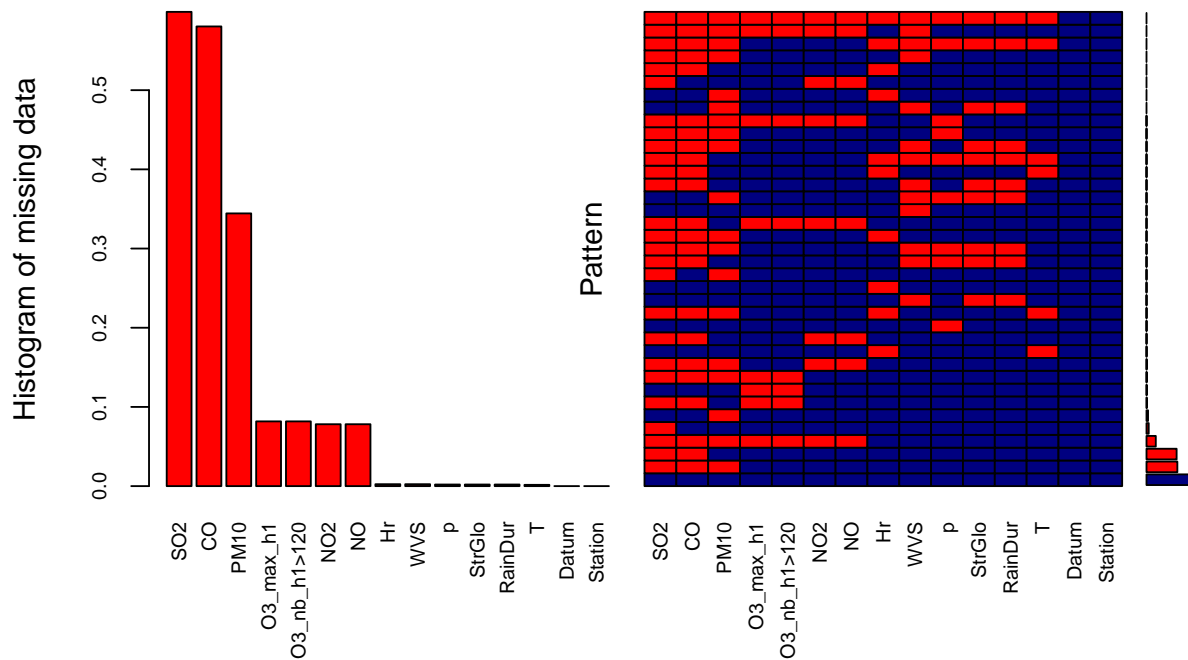
2.2 Aufgabe 2



2.3 Aufgabe 3

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 $50\mu\text{g}/\text{m}^3$ 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))

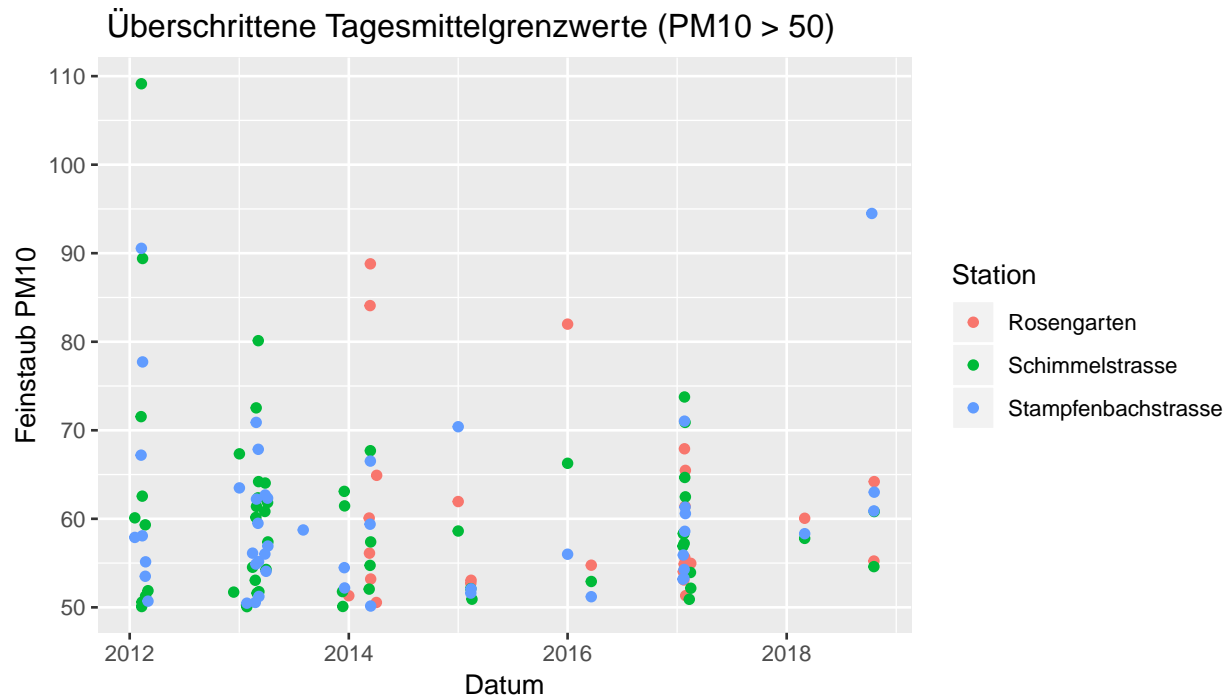
## # A tibble: 3 x 2
##   Station      n
##   <chr>    <int>
## 1 Rosengarten    25
## 2 Schimmelstrasse 56
## 3 Stampfenbachstrasse 47
```

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

2.4 Aufgabe 4

2.4.2 Überschrrittenen 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("Überschrrittenen Tagesmittelgrenzwerte (PM10 > 50)")
```



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 = $20 \mu\text{g}/\text{m}^3$

Daten vorbereiten: Nach dem Jahr und Station gruppieren

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

```
head(luftqual.PM10)
```

```
## # A tibble: 6 x 3  
## # Groups:   Jahr [2]  
##   Jahr Station      n  
##   <chr> <chr>    <int>  
## 1 2012 Rosengarten      0  
## 2 2012 Schimmelstrasse 11  
## 3 2012 Stampfenbachstrasse 8  
## 4 2013 Rosengarten      0
```

2.5 Aufgabe 5

```
## 5 2013 Schimmelstrasse      22
## 6 2013 Stampfenbachstrasse  19

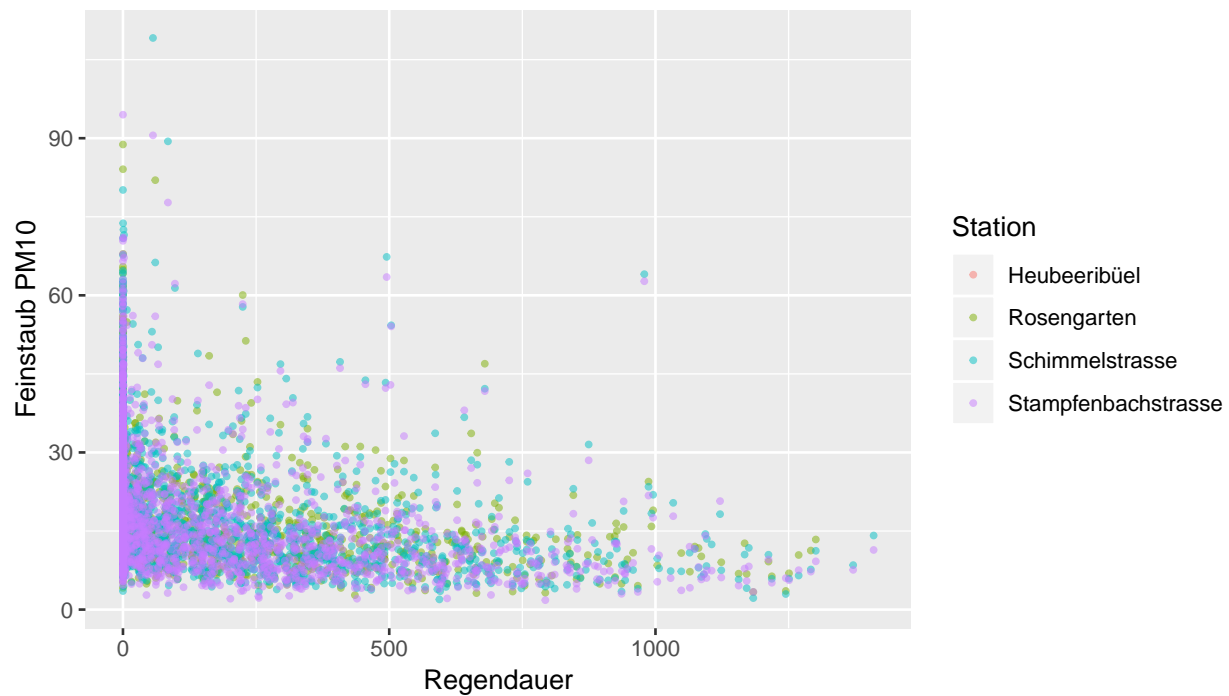
# pairwise wilcox test
# pairwise.wilcox.test(luftqual.PM10$n, luftqual.PM10$Jahr, luftqual.PM10$Station)
```

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.

2.6 Aufgabe 6

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.

$h_0: \mu_{\text{Feinstaub_regen}} = \mu_{\text{Feinstaub_keinregen}}$

$h_1: \mu_{\text{Feinstaub_regen}} < \mu_{\text{Feinstaub_keinregen}}$

```
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 == 0]
```

```
## 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
```

$p < 0.01 \rightarrow h_0$ 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