Hausaufgabe Wast3

J.Brändli, T.Haas, F.Kahlbacher

22 Dezember 2018

Inhaltsverzeichnis

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 (SO_2), Kohlenmonoxid (CO), Stickstoffdioxid (NO_2), Stickstoffmonoxid (NO), Ozonmengen (O_3) und meteorolgische Daten, sowie Feinstaubmessungen (PM10). Vor allem die Feinstaubmessungen sind von Relevanz. Der Feinstaub kann tief in die Lunge eindringen und schwerwiegende Auswirkungen auf die Gesunheit des Menschen haben. Die Messgrössen stammen von den Wetterstationen an der Stampfenbachstrasse, Schimmelstrasse, Heubeeribüel 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 unten) 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*	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

```
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, Heubeeribüel und Rosengartenstrasse) 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 für die anderen Stationen übernommen.

```
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.

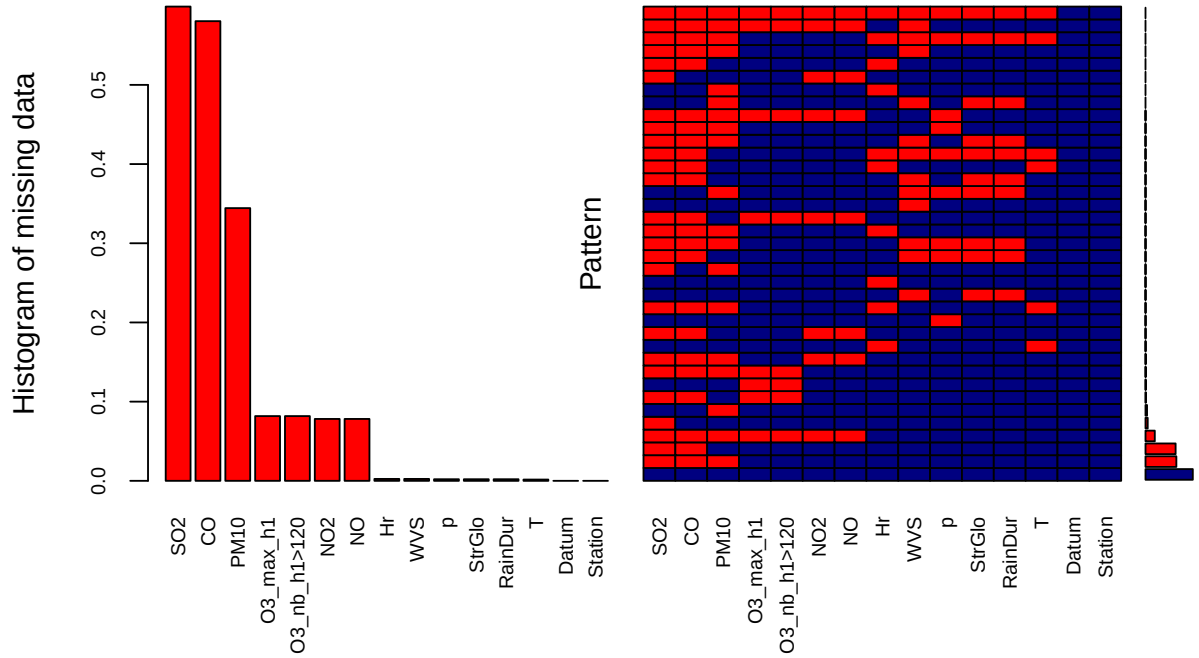
```
luftqual[2:14] <- as_tibble(sapply(luftqual[2:14], as.numeric))
```

2.2 Aufgabe 2

Die verschiedenen Messwerte verhalten sich über alle Stationen gesehen relativ ähnlich. Auffällig ist, dass die Messwerte nicht bei allen Stationen durchgängig vorhanden sind.

2.3 Aufgabe 3

Im Datensatz gibt es insgesamt 18528 fehlende Werte (NAs). 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 SO_2 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 SO_2, CO_2 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 von PM10 beträgt 50 $\frac{\mu g}{m^3}$ und darf max. 1 mal pro Jahr überschritten werden.

Die Daten werden wie folgt vorbereitet:

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

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

Datum	Feinstaub PM10	Station	PM10_uberschritt
2012-01-01	NaN	Rosengarten	NA
2012-01-02	NaN	Rosengarten	NA
2012-01-03	NaN	Rosengarten	NA
2012-01-04	NaN	Rosengarten	NA
2012-01-05	NaN	Rosengarten	NA
2012-01-06	NaN	Rosengarten	NA

Die ersten 6 Werte (siehe Tabelle oben) sind per Zufall gerade NaN respektiv NA.

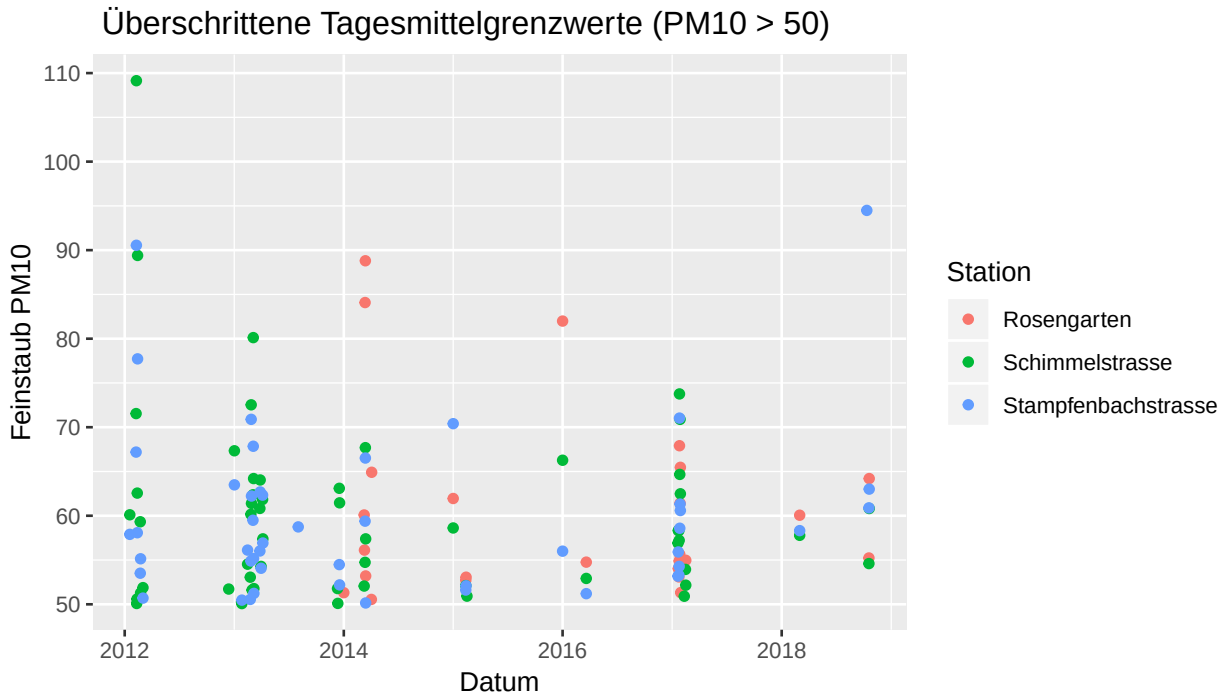
2.4.1 Wie oft wird der Tagesmittel-Grenzwert an welcher Station überschritten?

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

Station	n
Rosengarten	25
Schimmelstrasse	56
Stampfenbachstrasse	47

An der Schimmelstrasse wird der Tagesmittel-Grenzwert mit 56 -mal am häufigsten überschritten.

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



Der Tagesmittelgrenzwert wird bei allen Stationen oft in den gleichen Wochen / Monaten übertroffen. Dies ist gut möglich, 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-Überschreitungen signifikant grösser als zufällig

Der Jahresmittelgrenzwert von PM10 liegt bei 20 $\frac{\mu g}{m^3}$. Die Daten werden nach dem Jahr und der Station gruppiert und alle NAs werden entfernt.

```
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).

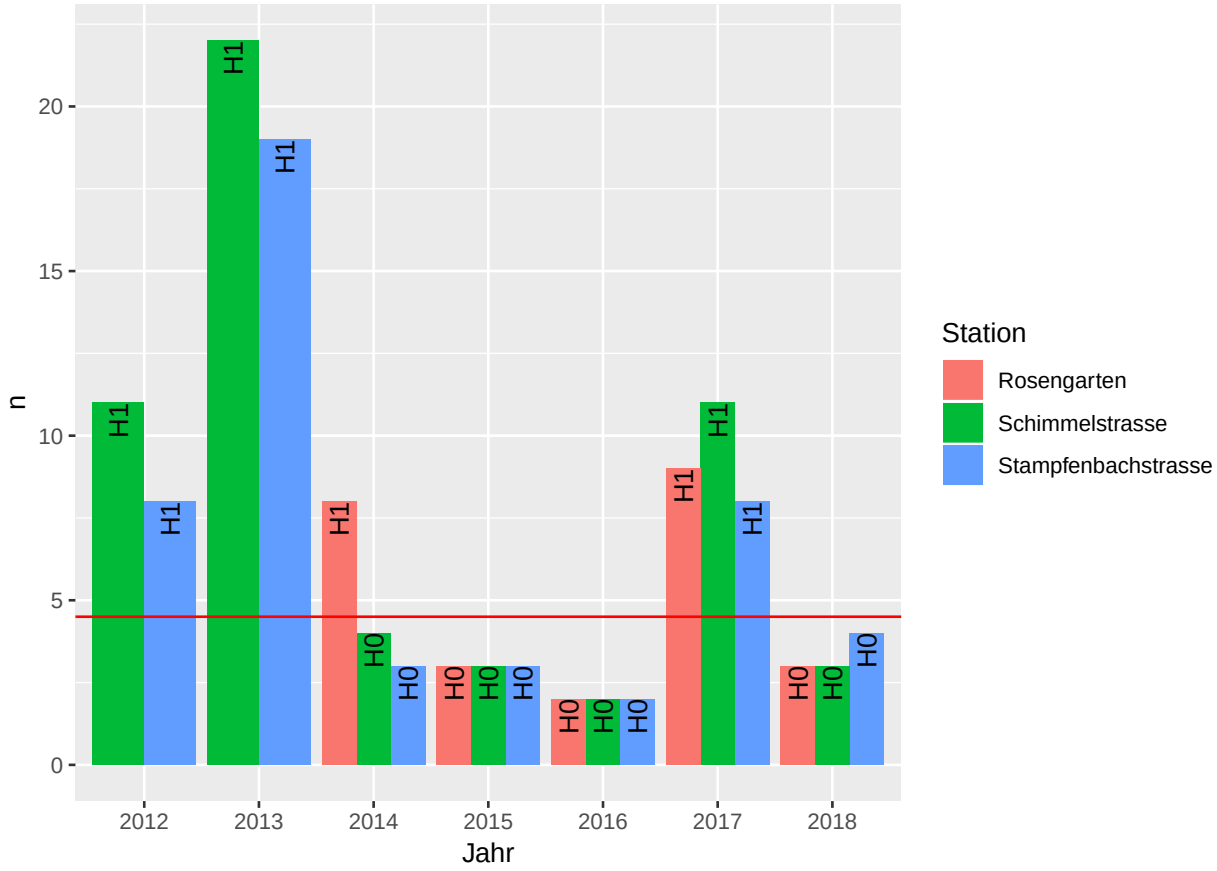
H_0: eine Überschreitung pro Jahr
H_1: mehr als 1 Überschreitung pro Jahr

```
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)
kable(luftqual.PM10.2)
```

Jahr	Station	n	h
2012	Schimmelstrasse	11	H1
2012	Stampfenbachstrasse	8	H1
2013	Schimmelstrasse	22	H1
2013	Stampfenbachstrasse	19	H1
2014	Rosengarten	8	H1
2014	Schimmelstrasse	4	H0
2014	Stampfenbachstrasse	3	H0
2015	Rosengarten	3	H0
2015	Schimmelstrasse	3	H0
2015	Stampfenbachstrasse	3	H0
2016	Rosengarten	2	H0
2016	Schimmelstrasse	2	H0
2016	Stampfenbachstrasse	2	H0
2017	Rosengarten	9	H1
2017	Schimmelstrasse	11	H1
2017	Stampfenbachstrasse	8	H1
2018	Rosengarten	3	H0
2018	Schimmelstrasse	3	H0
2018	Stampfenbachstrasse	4	H0

H_0 wird ab 5 Überschreitungen im Jahr verworfen.

Plot der Überschreitungen

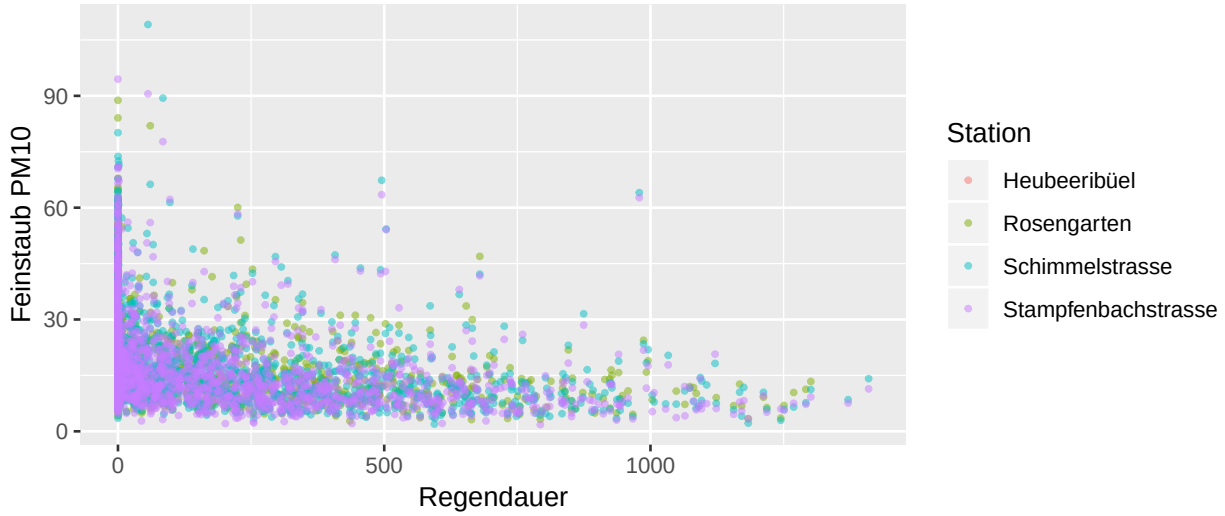


In den Jahren und Stationen bei welchen die Balken über der roten Linie liegen, wird H_0 verworfen. Bei allen darunter liegenden wird H_0 angenommen.

2.5 Aufgabe 5

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

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



Aus der Grafik ist folgendes Verhalten zu erkennen: Die Feinstaubkonzentration ist tendenziell höher wenn es nicht regnet, als wenn es regnet. Dieser Effekt nimmt mit der länge der Regendauer 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 ist.

H_0: mu_Feinstaub_regen = mu_Feinstaub_keinregen
H_1: 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
```

Der p-Wert (p-value) ist kleiner als $\alpha = 0.01$. Deshalb wird die Nullhypothese verworfen. Sie wird sogar klar verworfen. Das heisst, die Feinstaubkonzentration ist an Tagen mit Regen signifikant tiefer als an Tagen ohne Regen.

2.6 Aufgabe 6

Für diese Aufgabe werden drei neue Datensätze eingelesen. Diese enthalten Informationen über eine Bevökerungsbefragung der Stadt Zürich und über den Wohnort der Befragten.

```
bev_bef <- read_csv("bevoelkerungsbefragung_2015_stadtentwicklung_zuerich.csv")
adressen <- read_csv("adressen.csv")
attribute <- read_csv("attributbeschreibung_bvb_2015_stadtentwicklung_zuerich.csv")
```

Um die in der Aufgabenstellung vorgeschlagene Frage (Zufriedenheit mit dem öffentlichen Grünraum im Quartier) zu finden werden folgende Befehle verwendet:

```
filter(attribute, str_detect(feldbeschreibung, "öffentlichen Grünraum"))

attribute$feldbeschreibung[attribute$technischerfeldname == "f36105Lang"]
```

```
## # A tibble: 2 x 3
##   technischerfeldn~ sprechenderfeldname         feldbeschreibung
##   <chr>             <chr>                       <chr>
## 1 f36105Lang        F36105 Zufriedenheit mit d~ Zufriedenheit mit dem öff~
## 2 f36105Sort        F36105 Zufriedenheit mit d~ Zufriedenheit mit dem öff~
## [1] "Zufriedenheit mit dem öffentlichen Grünraum im Quartier (Fragebatterie)  [Beschreibung, String]
```

Dabei trifft die Frage *f36105* zu. Als nächstes wird gesucht in welchem Stadtkreis sich die jeweiligen Messstationen befinden.

```
filter(adressen, str_detect(lokalisationsname, "Stampfenbachstrasse"))
filter(adressen, str_detect(lokalisationsname, "Schimmelstrasse"))
filter(adressen, str_detect(lokalisationsname, "Rosengarten"))
```

Da die Messstation in Heubeeribüel keine Feinstaubmessungen durchgeführt haben, wurde diese nicht lokalisiert. Die Stampfenbachstrasse befindet sich im Kreis 1 und 6 welches dem Stadtkreiscode 1 und 5 enspricht. In Kreis 3 und 4 befindet sich die Schimmelstrasse (Stadkreiscode 2 und 3). Und zuletzt die Rosengartenstation welche sich im Kreis 10 (Stadtkreiscode 9) befindet. Der Stadtkreiscode weicht um von den Kreisen ab, da Kreis 1 und 2 zusammen einen Code haben.

In der Aufgabenstellung wird verlangt, dass die Antworten mit den Noten 1-3 zusammengefügt werden zur Note 3. Weiter werden alle Anworten mit der Note 98 (weiss nicht) und 99 (keine Angabe) entfernt.

```
# Noten 1-3 neu 3
bev_bef_gruen <- select(bev_bef,
                        intnr2015Sort, stadtkreiseLang, stadtkreiseSort, f36105Lang, f36105Sort)
bev_bef_gruen$f36105Sort[bev_bef_gruen$f36105Sort < 3] <- 3

# entfernen von weiss nicht 98 und keine Angabe 99
bev_bef_gruen <- bev_bef_gruen[!bev_bef_gruen$f36105Sort %in% c(98,99),]
```

Nun werden die Noten der Kreise der Messstation zugewiesen.

```
kreis_stampfen <- filter(bev_bef_gruen, stadtkreiseSort %in% c(1,5))
kreis_schimmel <- filter(bev_bef_gruen, stadtkreiseSort %in% c(2,3))
kreis_rosen <- filter(bev_bef_gruen, stadtkreiseSort == 9)
```

Um sich eine Übersicht über die Feinstaubbelastung und Zufriedenheit der Bevölkerung nach Station zu verschaffen werden die jeweiligen Mittelwerte berechnet und in untenstehenden Tabelle dargestellt.

Station	Mittelwert.Noten	Mittelwert.PM10
Rosengarten	5.458875	18.71451
Schimmelstrasse	5.069098	19.23306
Stampfenbachstrasse	5.389022	18.16022

Hierbei ist erkennbar das an der Schimmelstrasse eine höhere durchschnittliche Feinstaubbelastung gemessen wurde. Zugleich ist auch die durchschnittliche Zufriedenheit der Bevölkerung tiefer als bei den anderen Messstationen.

Um festzustellen, ob die Zufriedenheit nur zufällig tiefer ist, wird getestet. Dabei ist die Nullhypthese, dass es keinen Zufriedenheitsunterschied zwischen den Stadtkreisen gibt.

```
wilcox.test(x = kreis_rosen$f36105Sort, y = kreis_schimmel$f36105Sort,
       conf.level = 0.99, alternative = "two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  kreis_rosen$f36105Sort and kreis_schimmel$f36105Sort
## W = 74963, p-value = 7.262e-09
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(x = kreis_rosen$f36105Sort, y = kreis_stampfen$f36105Sort,
       conf.level = 0.99, alternative = "two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  kreis_rosen$f36105Sort and kreis_stampfen$f36105Sort
## W = 50698, p-value = 0.2561
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(x = kreis_schimmel$f36105Sort, y = kreis_stampfen$f36105Sort,
       conf.level = 0.99, alternative = "two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  kreis_schimmel$f36105Sort and kreis_stampfen$f36105Sort
## W = 87214, p-value = 1.066e-08
## alternative hypothesis: true location shift is not equal to 0
```

Die Nullhypothese wird beim ersten und dritten Test verworfen. Das heisst, dass die Zufriedenheit der Bevölkerung im Kreis 3 und 4 signifikant tiefer ist.

3 Fazit

In dieser Aufgabe wurden Luftqualitätsmessungen und Bevölkerungsbefragungen in der Stadt Zürich ausgewertet. Aus unseren Untersuchungen geht hervor, dass die Regendauer einen Einfluss auf die Feinstaubkonzentration in der Luft hat. Ebenfalls sind die Personen, welche im Stadtkreis 3 und 4 Wohnhaft sind unzufriedener mit dem öffentlichen Grünraum im Quartier, als in den Stadkreisen 1, 2, 5 und 10. Die Schwierigkeit dieser Aufgabe bestand in der Datenaufbereitung, welche auch am meisten Zeit in Anspruch genommen hat.