DSP_Datenanalyse

Melanie Weissenboeck

2023-01-08

```
library(xlsx)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.4.0     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(vcd)
```

```
## Lade nötiges Paket: grid
```

Laden der Datensets

Die gegebenen Daten werden aus den Excel-Files eingelesen und als Dataframes gespeichert. Für die verschiedenen Datensets werden die Frames mit `ds1`, `ds2` und `ds3` bezeichnet. Zur besseren Übersicht werden von einem Datenset die ersten fünf Zeilen ausgegeben.

```
head(ds1)
```

```
##   ANF_ID                          ANF_NAME
## 1  57209 Berichtigungsprotokoll Zulassung
## 2  55910                    Indexklasse(n)
## 3  55904    Reindizierung der Indexklasse
## 4  55639                      eVtg Polizze
## 5  55643                      eVtg Polizze
## 6  55641                      eVtg Polizze
##
## 1
## 2
## 3 Nach Reindizierung der Indexklasse wird der Eintrag in der Trefferliste upgedatet (z.B.: Statt "An
## 4
## 5
## 6
##   ANF_FEHLERWAHRSCHEINLICHKEIT ANF_FEHLERKOSTEN ANF_RISIKO TF_ID
## 1                         <NA>             <NA>       <NA> 41104
## 2                         hoch             hoch       hoch 40794
## 3                      gering              hoch     mittel 40794
## 4                      gering           gering     gering 40795
```

```
## 5                       gering           gering     gering 40796
## 6                       gering           gering     gering 40797
##                                            TF_NAME TF_ABDECKUNG
## 1            Zeus_Berichtigungsprotokoll Zulassung        100.0
## 2 GLI_DF_Reindizieren_Idx_auf_Partnerkorrespondenz         16.6
## 3 GLI_DF_Reindizieren_Idx_auf_Partnerkorrespondenz         16.6
## 4                    GLI_DF_Formatcheck_eVtg_PolNr        100.0
## 5               GLI_MENÜ_IV_Formatcheck_eVtg_PolNr        100.0
## 6                    GLI_SF_Formatcheck_eVtg_PolNr        100.0
##
## 1 Folgende eFlow Services müssen auf dem Server at010000sat34 als Autorun Stations gestartet sein\r\n
## 2
## 3
## 4
## 5
## 6
##
## 1 1 FilePortal\r\n- eFlow Module Activator starten\r\n- Typ "Zeus" auswählen\r\n- Symbol "FilePortal"
## 2
## 3
## 4
## 5
## 6
##
## 1 - Das Dokument ist in ARC vorhanden\r\n- Kontrolle der Felder:\r\n  * Ausstellungsdatum\r\n  * Zula
## 2          - Die Indexklasse des ausgewählten Eintrages in der Trefferliste ändert sich auf "Pa.Korr"
## 3          - Die Indexklasse des ausgewählten Eintrages in der Trefferliste ändert sich auf "Pa.Korr"
## 4
## 5
## 6
##   RES1_STATUS  RES1_RELEASE RES2_STATUS  RES2_RELEASE RES3_STATUS  RES3_RELEASE
## 1          OK Release 20.10        <NA>          <NA>        <NA>          <NA>
## 2          OK Release 21.30          OK Release 21.30          OK Release 21.30
## 3          OK Release 21.30          OK Release 21.30          OK Release 21.30
## 4          OK Release 21.30          OK Release 21.30          OK Release 21.30
## 5          OK Release 21.30          OK Release 21.30          OK Release 21.30
## 6          OK Release 21.30          OK Release 21.30          OK Release 21.30
```

Die Spalte `ANF_BESCHREIBUNG` dient als Prädiktor für die Klassifikation der Spalte `ANF_RISIKO`. Daher wird im ersten Schritt der Analyse überprüft, wieviele Datensätze in den jeweiligen Sets mit diesen beiden Werten vorhanden sind:

```
# Anzahl der Datensaetze mit AF Beschreibung und Risiko
touse_ds1 <- nrow(ds1) - sum(is.na(ds1$ANF_BESCHREIBUNG) | is.na(ds1$ANF_RISIKO))
touse_ds2 <- nrow(ds2) - sum(is.na(ds2$ANF_BESCHREIBUNG) | is.na(ds2$ANF_RISIKO))
touse_ds3 <- nrow(ds3) - sum(is.na(ds3$ANF_BESCHREIBUNG) | is.na(ds3$ANF_RISIKO))
print(paste0("ds1: ", touse_ds1))
```

```
## [1] "ds1: 383"
```

```
print(paste0("ds2: ", touse_ds2))
```

```
## [1] "ds2: 3142"
```

```
print(paste0("ds3: ", touse_ds3))
```

```
## [1] "ds3: 1485"
```

Datenaufbereitung

Jene Datensätze, wo entweder `ANF_BESCHREIBUNG` oder `ANF_RISIKO` fehlt, können für die Klassifikation nicht verwendet werden, und werden daher aus den Datensets entfernt.

```
# Datenset darf nur Zeilen enthalten mit ANF_RISIKO und ANF_BESCHREIBUNG
ds1 <- ds1[complete.cases(ds1[ , c(3,6)]),]
ds2 <- ds2[complete.cases(ds2[ , c(3,6)]),]
ds3 <- ds3[complete.cases(ds3[ , c(3,6)]),]
```

Im nächsten Schritt werden ordinale Variablen als Faktoren definiert und deren Ausprägungen sortiert.

```
# Faktoren definieren
ds1$ANF_FEHLERWAHRSCHEINLICHKEIT <- as.factor(ds1$ANF_FEHLERWAHRSCHEINLICHKEIT)
ds1$ANF_FEHLERKOSTEN <- as.factor(ds1$ANF_FEHLERKOSTEN)
ds1$ANF_RISIKO <- as.factor(ds1$ANF_RISIKO)
ds1$ANF_RISIKO <- ordered(ds1$ANF_RISIKO, levels = c("gering", "mittel", "hoch"))
ds1$ANF_FEHLERKOSTEN <- ordered(ds1$ANF_FEHLERKOSTEN, levels = c("gering", "mittel", "hoch"))
ds1$ANF_FEHLERWAHRSCHEINLICHKEIT <- ordered(ds1$ANF_FEHLERWAHRSCHEINLICHKEIT, levels = c("gering", "mit

ds2$ANF_FEHLERWAHRSCHEINLICHKEIT <- as.factor(ds2$ANF_FEHLERWAHRSCHEINLICHKEIT)
ds2$ANF_FEHLERKOSTEN <- as.factor(ds2$ANF_FEHLERKOSTEN)
ds2$ANF_RISIKO <- as.factor(ds2$ANF_RISIKO)
ds2$ANF_RISIKO <- ordered(ds2$ANF_RISIKO, levels = c("gering", "mittel", "hoch"))
ds2$ANF_FEHLERKOSTEN <- ordered(ds2$ANF_FEHLERKOSTEN, levels = c("gering", "mittel", "hoch"))
ds2$ANF_FEHLERWAHRSCHEINLICHKEIT <- ordered(ds2$ANF_FEHLERWAHRSCHEINLICHKEIT, levels = c("gering", "mit

ds3$ANF_FEHLERWAHRSCHEINLICHKEIT <- as.factor(ds3$ANF_FEHLERWAHRSCHEINLICHKEIT)
ds3$ANF_FEHLERKOSTEN <- as.factor(ds3$ANF_FEHLERKOSTEN)
ds3$ANF_RISIKO <- as.factor(ds3$ANF_RISIKO)
ds3$ANF_RISIKO <- ordered(ds3$ANF_RISIKO, levels = c("gering", "mittel", "hoch"))
ds3$ANF_FEHLERKOSTEN <- ordered(ds3$ANF_FEHLERKOSTEN, levels = c("gering", "mittel", "hoch"))
ds3$ANF_FEHLERWAHRSCHEINLICHKEIT <- ordered(ds3$ANF_FEHLERWAHRSCHEINLICHKEIT, levels = c("gering", "mit
```

In den Datensets gibt es unterschiedliche Angaben in den Spalten `RES1...` und `RES2...`. Diese verschiedenen Werte werden im Folgenden in die neuen Variablen `AKT_RES_STATUS` und `AKT_RES_RELEASE` transformiert. Diese beiden Spalten werden anschließend wieder als Faktor dargestellt und einige Ausprägungen zusammengefasst.

```
# Berechnung aktuellstes Resultat fuer ds1
for (i in (1:nrow(ds1))){
  if ((is.na(ds1$RES3_STATUS[i]))==FALSE){
    ds1$AKT_RES_STATUS[i] <- ds1$RES3_STATUS[i]
    ds1$AKT_RES_RELEASE[i] <- ds1$RES3_RELEASE[i]
  }

  if ((is.na(ds1$RES3_STATUS[i])==TRUE) & (is.na(ds1$RES2_STATUS[i]))==FALSE){
      ds1$AKT_RES_STATUS[i] <- ds1$RES2_STATUS[i]
      ds1$AKT_RES_RELEASE[i] <- ds1$RES2_RELEASE[i]
  }

  if ((is.na(ds1$RES3_STATUS[i])==TRUE) & (is.na(ds1$RES2_STATUS[i])==TRUE)){
      ds1$AKT_RES_STATUS[i] <- ds1$RES1_STATUS[i]
      ds1$AKT_RES_RELEASE[i] <- ds1$RES1_RELEASE[i]
  }
}
```

```r
ds1 <- ds1[, -c(13,14,15,16,17,18)]
```

```r
# Berechnung aktuellstes Resultat fuer ds2
for (i in (1:nrow(ds2))){
  if ((is.na(ds2$RES3_STATUS[i]))==FALSE){
    ds2$AKT_RES_STATUS[i] <- ds2$RES3_STATUS[i]
    ds2$AKT_RES_RELEASE[i] <- ds2$RES3_RELEASE[i]
  }

  if ((is.na(ds2$RES3_STATUS[i])==TRUE) & (is.na(ds2$RES2_STATUS[i]))==FALSE){
      ds2$AKT_RES_STATUS[i] <- ds2$RES2_STATUS[i]
      ds2$AKT_RES_RELEASE[i] <- ds2$RES2_RELEASE[i]
  }

  if ((is.na(ds2$RES3_STATUS[i])==TRUE) & (is.na(ds2$RES2_STATUS[i])==TRUE)){
      ds2$AKT_RES_STATUS[i] <- ds2$RES1_STATUS[i]
      ds2$AKT_RES_RELEASE[i] <- ds2$RES1_RELEASE[i]
  }
}
ds2 <- ds2[, -c(13,14,15,16,17,18)]
```

```r
# Berechnung aktuellstes Resultat fuer ds3
for (i in (1:nrow(ds3))){
  if ((is.na(ds3$RES3_STATUS[i]))==FALSE){
    ds3$AKT_RES_STATUS[i] <- ds3$RES3_STATUS[i]
    ds3$AKT_RES_RELEASE[i] <- ds3$RES3_RELEASE[i]
  }

  if ((is.na(ds3$RES3_STATUS[i])==TRUE) & (is.na(ds3$RES2_STATUS[i]))==FALSE){
      ds3$AKT_RES_STATUS[i] <- ds3$RES2_STATUS[i]
      ds3$AKT_RES_RELEASE[i] <- ds3$RES2_RELEASE[i]
  }

  if ((is.na(ds3$RES3_STATUS[i])==TRUE) & (is.na(ds3$RES2_STATUS[i])==TRUE)){
      ds3$AKT_RES_STATUS[i] <- ds3$RES1_STATUS[i]
      ds3$AKT_RES_RELEASE[i] <- ds3$RES1_RELEASE[i]
  }
}
ds3 <- ds3[, -c(13,14,15,16,17,18)]
```

```r
# neue Resultat Spalten als Faktor
ds1$AKT_RES_RELEASE <- as.factor(ds1$AKT_RES_RELEASE)
ds1$AKT_RES_STATUS <- as.factor(ds1$AKT_RES_STATUS)

ds2$AKT_RES_RELEASE <- as.factor(ds2$AKT_RES_RELEASE)
ds2$AKT_RES_STATUS <- as.factor(ds2$AKT_RES_STATUS)

ds3$AKT_RES_RELEASE <- as.factor(ds3$AKT_RES_RELEASE)
ds3$AKT_RES_STATUS <- as.factor(ds3$AKT_RES_STATUS)
```

```r
# Levels Status anzeigen
levels(ds1$AKT_RES_STATUS)
```

```
## [1] "FAILED"      "FAILED_GATE" "OK"          "OPEN"
```

```r
levels(ds2$AKT_RES_STATUS)
```

```
## [1] "FAILED"      "FAILED_GATE" "OK"          "OPEN"        "SKIPPED"
## [6] "TOUCHED"
```

```r
levels(ds3$AKT_RES_STATUS)
```

```
## [1] "FAILED"      "FAILED_GATE" "OK"          "OPEN"
```

```r
# Levels fuer ds1 zusammenfassen
levels(ds1$AKT_RES_STATUS) <- list(FAILED  = "FAILED_GATE", FAILED = "FAILED", OK = "OK", OPEN = "OPEN")
```

```r
# Levels fuer ds2 zusammenfassen
levels(ds2$AKT_RES_STATUS) <- list(FAILED  = "FAILED_GATE", FAILED = "FAILED", OK = "OK", OPEN = "OPEN"
```

```r
# Levels fuer ds3 zusammenfassen
levels(ds3$AKT_RES_STATUS) <- list(FAILED  = "FAILED_GATE", FAILED = "FAILED", OK = "OK", OPEN = "OPEN")
```

```r
# Levels Release anzeigen
levels(ds1$AKT_RES_RELEASE)
```

```
## [1] "Release 17.20" "Release 17.30" "Release 17.40" "Release 20.20"
## [5] "Release 21.30" "Release 21.40" "Release 22.10" "Release 22.20"
## [9] "Release 22.30"
```

```r
levels(ds2$AKT_RES_RELEASE)
```

```
##  [1] "Release 12.30" "Release 16.10" "Release 17.20" "Release 18.40"
##  [5] "Release 19.20" "Release 19.30" "Release 20.20" "Release 20.30"
##  [9] "Release 20.40" "Release 21.10" "Release 21.20" "Release 21.30"
## [13] "Release 21.40" "Release 22.10" "Release 22.20" "Release 22.30"
```

```r
levels(ds3$AKT_RES_RELEASE)
```

```
## [1] "Release 16.40" "Release 18.10" "Release 21.40" "Release 22.10"
## [5] "Release 22.20" "Release 22.30"
```

```r
# Levels fuer ds1 zusammenfassen
levels(ds1$AKT_RES_RELEASE) <- list(
  OLDERT21 = "Release 17.20",
  OLDERT21 = "Release 17.30",
  OLDERT21 = "Release 17.40",
  OLDERT21 = "Release 20.20",
  "21x" = "Release 21.30",
  "21x" = "Release 21.40",
  "22.10" = "Release 22.10",
  "22.20" = "Release 22.20",
  "22.30" = "Release 22.30")
```

```r
# Levels fuer ds2 zusammenfassen
levels(ds2$AKT_RES_RELEASE) <- list(
  OLDERT21 = "Release 12.30",
  OLDERT21 = "Release 16.10",
  OLDERT21 = "Release 17.20",
  OLDERT21 = "Release 18.40",
  OLDERT21 = "Release 19.20",
  OLDERT21 = "Release 19.30",
  OLDERT21 = "Release 20.20",
  OLDERT21 = "Release 20.30",
```

```
  OLDERT21 = "Release 20.40",
  "21x" = "Release 21.10",
  "21x" = "Release 21.20",
  "21x" = "Release 21.30",
  "21x" = "Release 21.40",
  "22.10" = "Release 22.10",
  "22.20" = "Release 22.20",
  "22.30" = "Release 22.30")
```

```
# Levels fuer ds3 zusammenfassen
levels(ds3$AKT_RES_RELEASE) <- list(
  OLDERT21 = "Release 16.40",
  OLDERT21 = "Release 18.10",
  "21x" = "Release 21.40",
  "22.10" = "Release 22.10",
  "22.20" = "Release 22.20",
  "22.30" = "Release 22.30")
```

```
# nicht benoetigte Spalten entfernen
ds1 <- ds1[, -c(7,8,10,11,12)]
ds2 <- ds2[, -c(7,8,10,11,12)]
ds3 <- ds3[, -c(7,8,10,11,12)]
```

```
summary(ds1)
```

```
##     ANF_ID             ANF_NAME         ANF_BESCHREIBUNG
##  Length:383         Length:383         Length:383
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  ANF_FEHLERWAHRSCHEINLICHKEIT ANF_FEHLERKOSTEN  ANF_RISIKO   TF_ABDECKUNG
##  gering:182                   gering:149       gering:158   Min.   :  0.00
##  mittel: 66                   mittel: 12       mittel:141   1st Qu.: 50.00
##  hoch  : 51                   hoch  :138       hoch  : 84   Median :100.00
##  NA's  : 84                   NA's  : 84                    Mean   : 80.16
##                                                             3rd Qu.:100.00
##                                                             Max.   :100.00
##  AKT_RES_STATUS AKT_RES_RELEASE
##  FAILED: 12     OLDERT21: 84
##  OK    :359     21x     : 30
##  OPEN  :  7     22.10   :132
##  NA's  :  5     22.20   :129
##                 22.30   :  3
##                 NA's    :  5
```

Da im `ds1` die Spalten für die Fehlerkosten und -wahrscheinlichkeit sehr viele `NAs` enthalten, werden diese beiden Variablen entfernt.

```
# Spalten FEHLERKOSTEN und FEHLERWS aus ds1 streichen,
# da zu viele NAs zur Gesamtzahl an Zeilen
ds1 <- ds1[, -c(4,5)]
ds1 <- ds1[complete.cases(ds1[, c(6,7)]),]
```

```
summary(ds2)
```

```
##     ANF_ID             ANF_NAME         ANF_BESCHREIBUNG
##  Length:3142        Length:3142        Length:3142
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  ANF_FEHLERWAHRSCHEINLICHKEIT ANF_FEHLERKOSTEN  ANF_RISIKO    TF_ABDECKUNG
##  gering:1002                  gering: 716      gering: 637   Min.   : -0.70
##  mittel:1310                  mittel: 858      mittel:1378   1st Qu.:  3.45
##  hoch  : 829                  hoch  :1567      hoch  :1127   Median : 16.50
##  NA's  :   1                  NA's  :   1                    Mean   : 29.68
##                                                              3rd Qu.: 50.00
##                                                              Max.   :100.00
##  AKT_RES_STATUS AKT_RES_RELEASE
##  FAILED: 577    OLDERT21: 489
##  OK    :2364    21x     :1146
##  OPEN  : 181    22.10   : 434
##  NA's  :  20    22.20   : 727
##                 22.30   : 326
##                 NA's    :  20
```

```
ds2 <- ds2[complete.cases(ds2[, c(4,5,8,9)]),]
```

```
summary(ds3)
```

```
##     ANF_ID             ANF_NAME         ANF_BESCHREIBUNG
##  Length:1485        Length:1485        Length:1485
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  ANF_FEHLERWAHRSCHEINLICHKEIT ANF_FEHLERKOSTEN  ANF_RISIKO   TF_ABDECKUNG
##  gering:233                   gering:178       gering:241   Min.   :  0.00
##  mittel:966                   mittel:804       mittel:670   1st Qu.:  7.10
##  hoch  :247                   hoch  :464       hoch  :574   Median : 14.30
##  NA's  : 39                   NA's  : 39                    Mean   : 25.22
##                                                             3rd Qu.: 33.33
##                                                             Max.   :100.00
##  AKT_RES_STATUS AKT_RES_RELEASE
##  FAILED:  49    OLDERT21:   4
##  OK    :1433    21x     :  48
##  OPEN  :   3    22.10   :1119
##                 22.20   :  12
##                 22.30   : 302
##
```

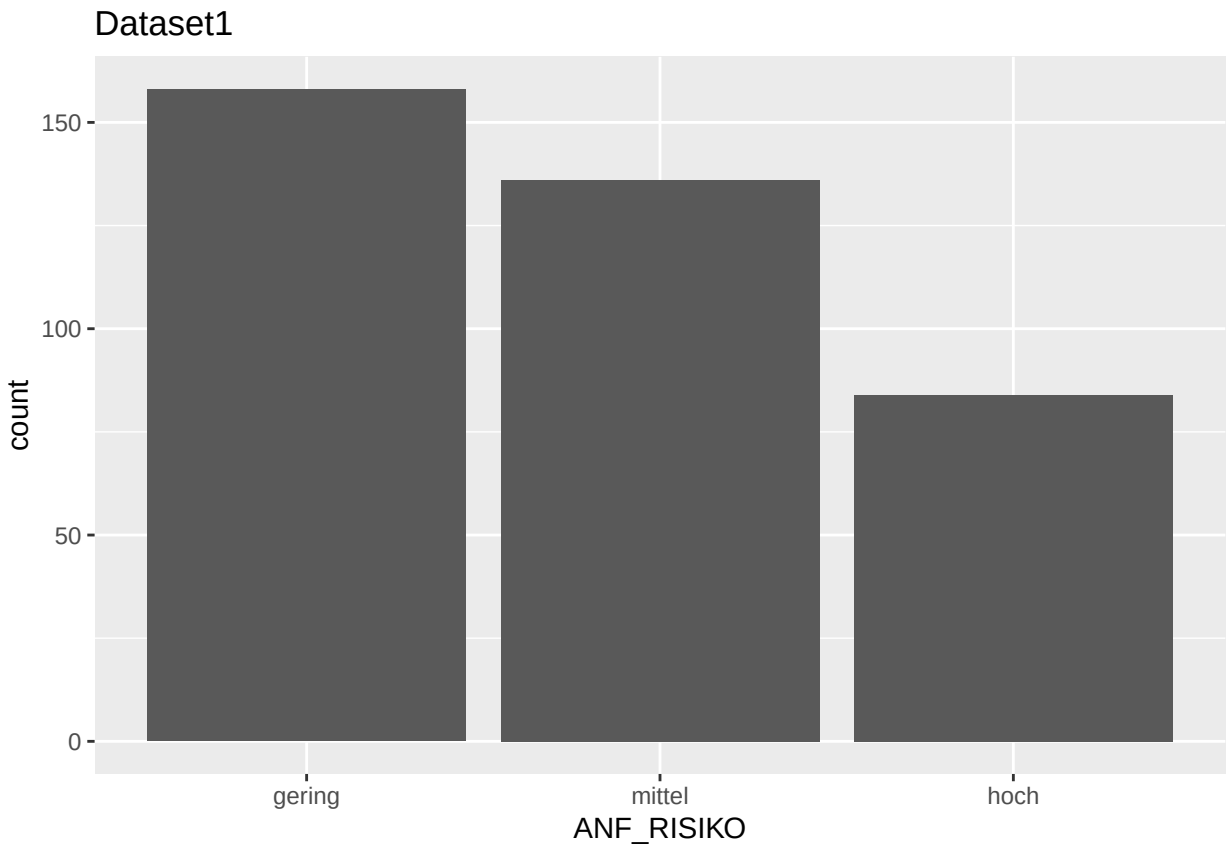
```
ds3 <- ds3[complete.cases(ds3[, c(4,5)]),]
```

Datenanalyse

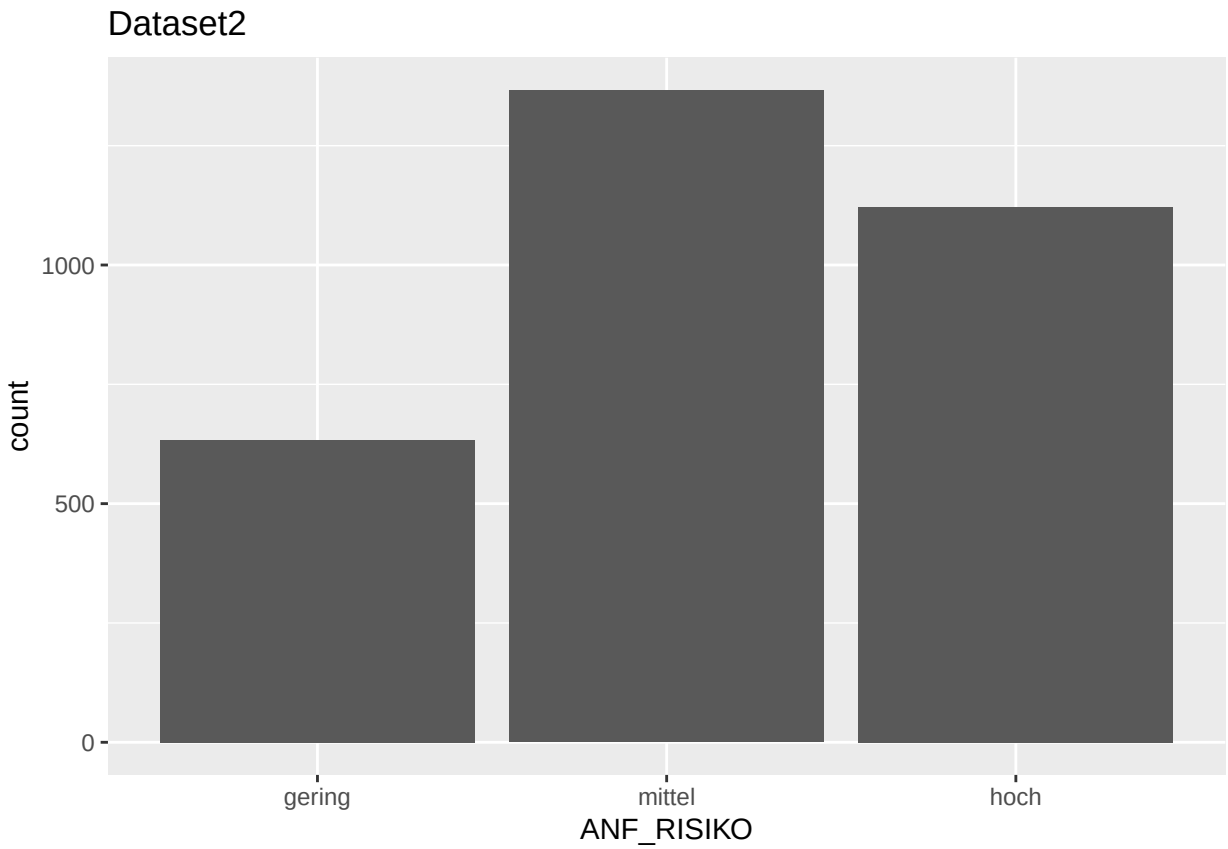
kategoriale Variablen

Um die Balance der Datensets anhand der Zielvariablen zu untersuchen, werden die absoluten Häufigkeiten in Balkendiagrammen dargestellt.

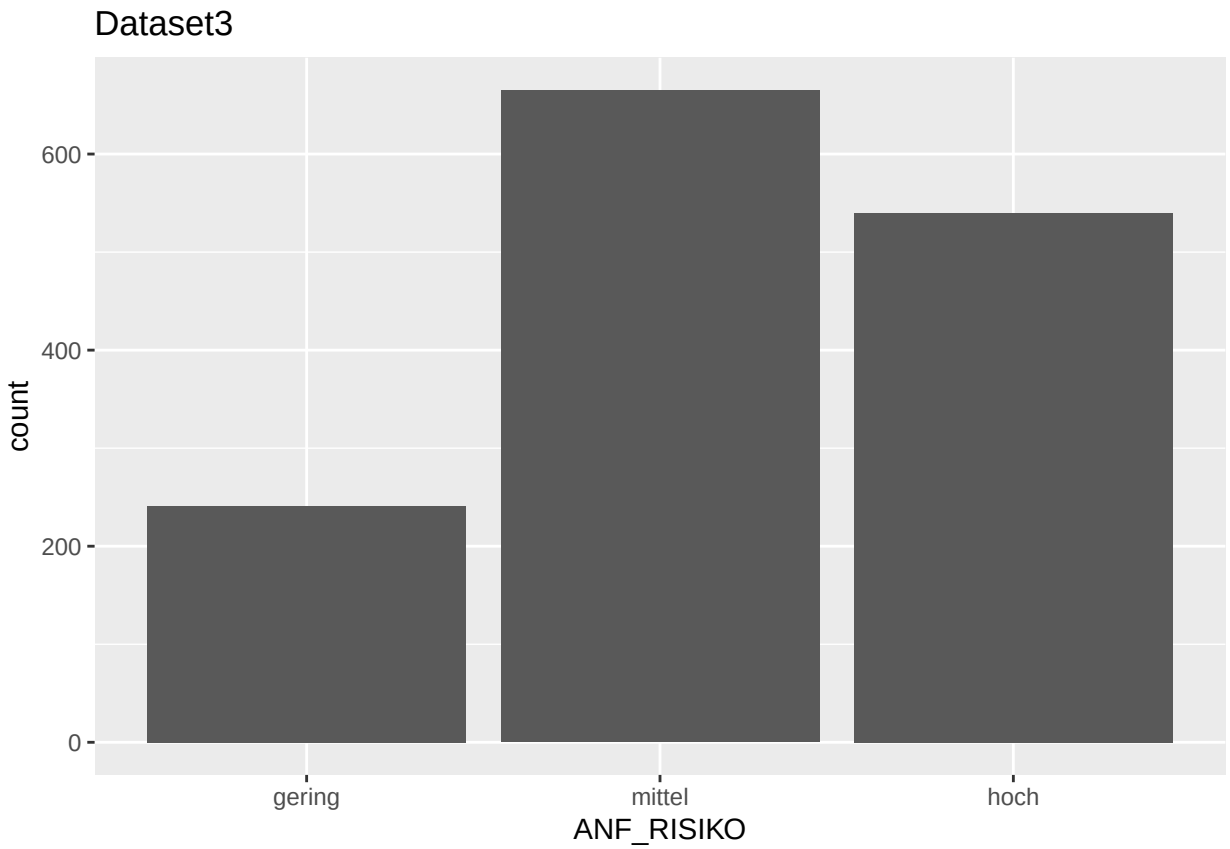
```
# absolute Haeufigkeiten der RISIKO Levels
par(mfrow=c(1,3))
ggplot(data = ds1) + geom_bar(mapping = aes(x = ANF_RISIKO)) + ggtitle("Dataset1")
```




```
ggplot(data = ds2) + geom_bar(mapping = aes(x = ANF_RISIKO)) + ggtitle("Dataset2")
```

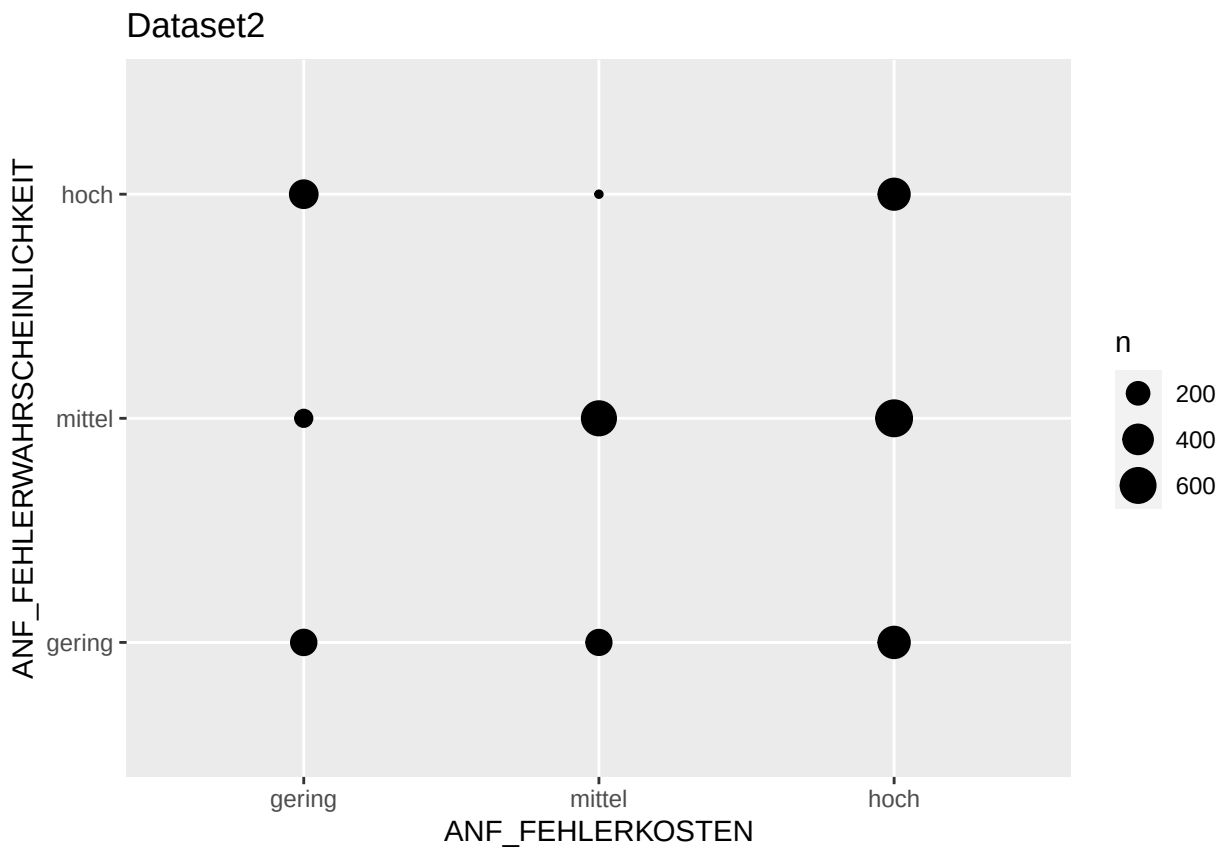



```
ggplot(data = ds3) + geom_bar(mapping = aes(x = ANF_RISIKO)) + ggtitle("Dataset3")
```

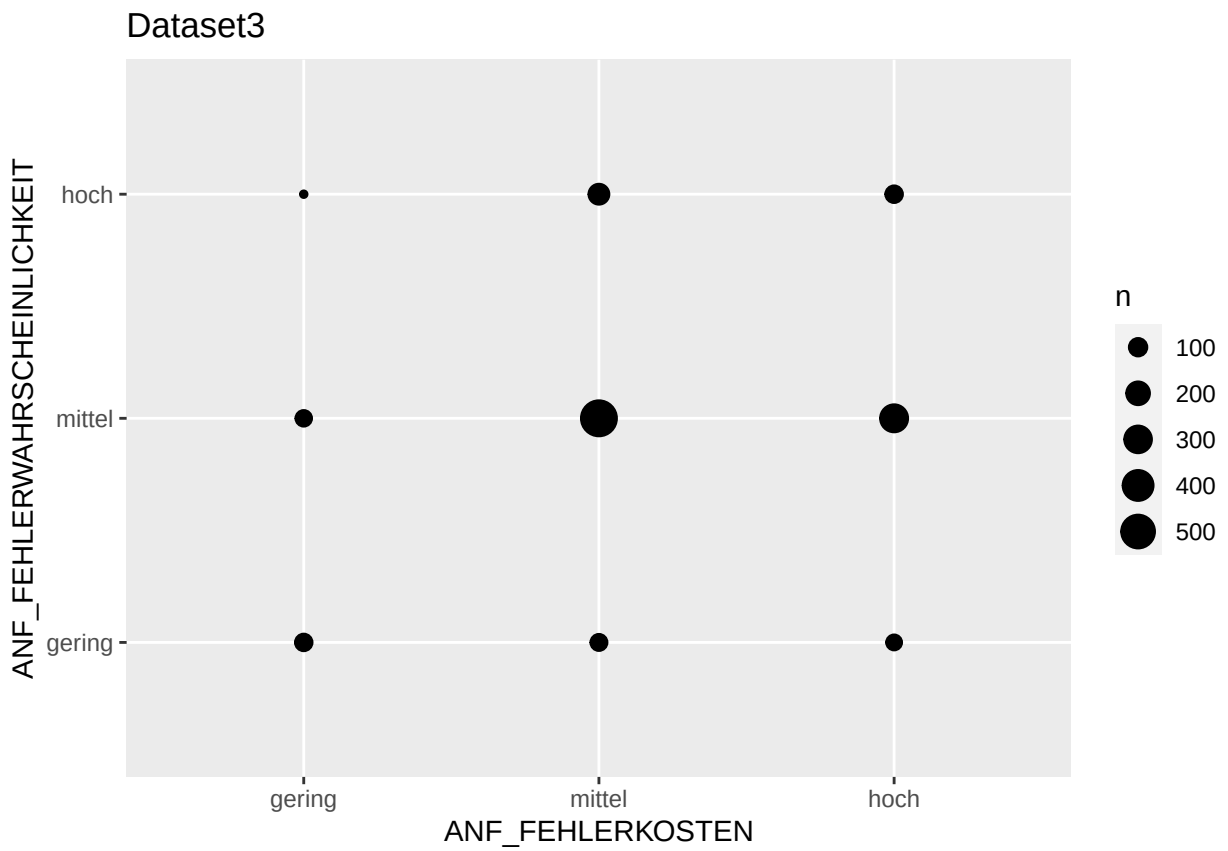



```
par(mfrow=c(1,1))
```

```
# absolute Haeufigkeiten von FEHLERKOSTEN und FEHLER-WS nach RISIKO
par(mfrow=c(1,2))
ggplot(data = ds2) + geom_count(mapping = aes(x = ANF_FEHLERKOSTEN, y = ANF_FEHLERWAHRSCHEINLICHKEIT))+
```

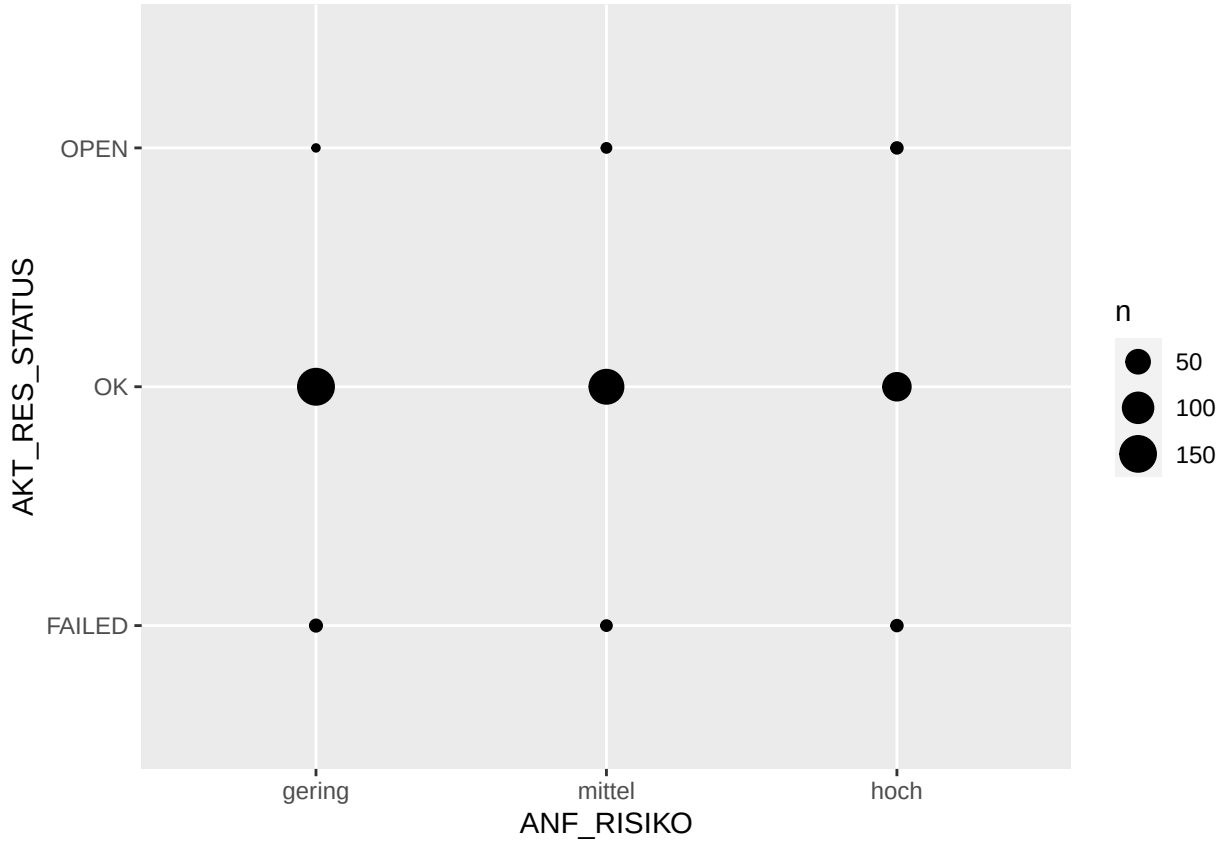



```
ggplot(data = ds3) + geom_count(mapping = aes(x = ANF_FEHLERKOSTEN, y = ANF_FEHLERWAHRSCHEINLICHKEIT))+
```

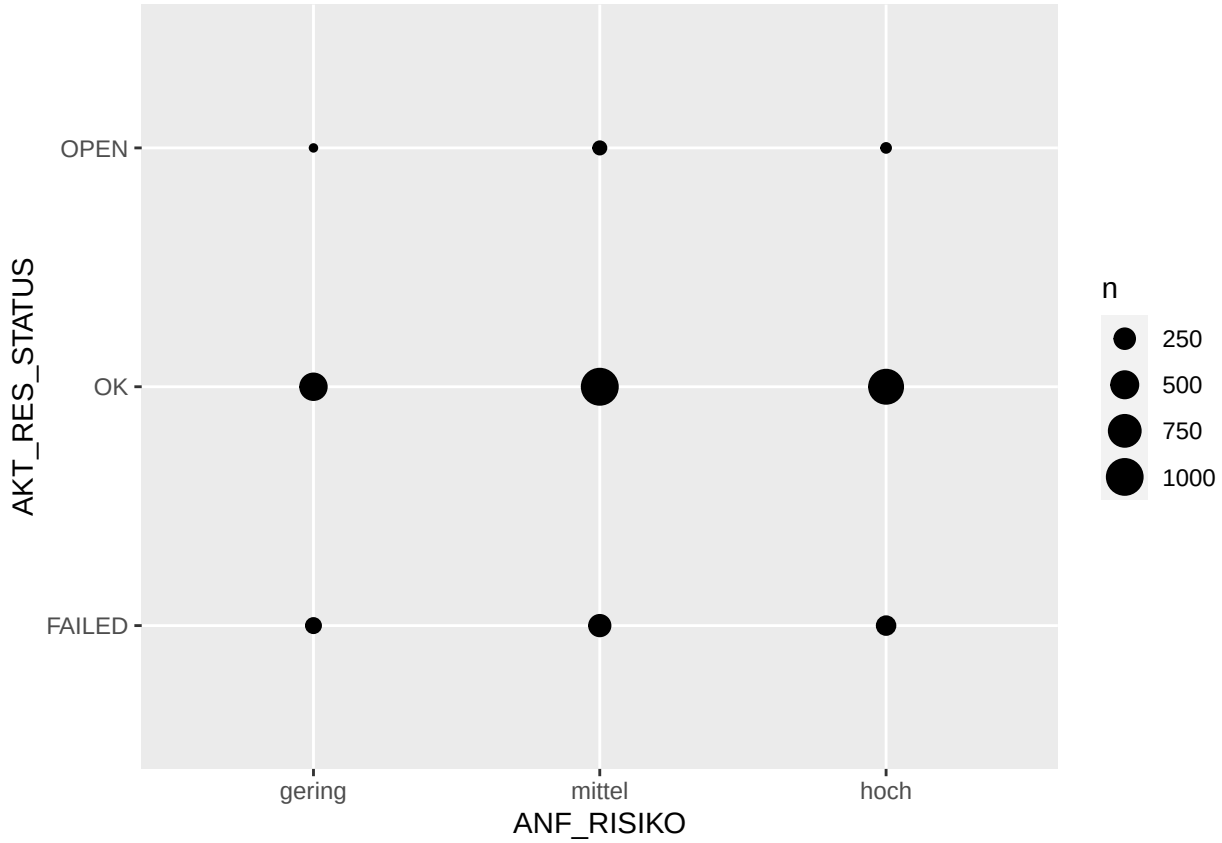



```
par(mfrow=c(1,1))
```

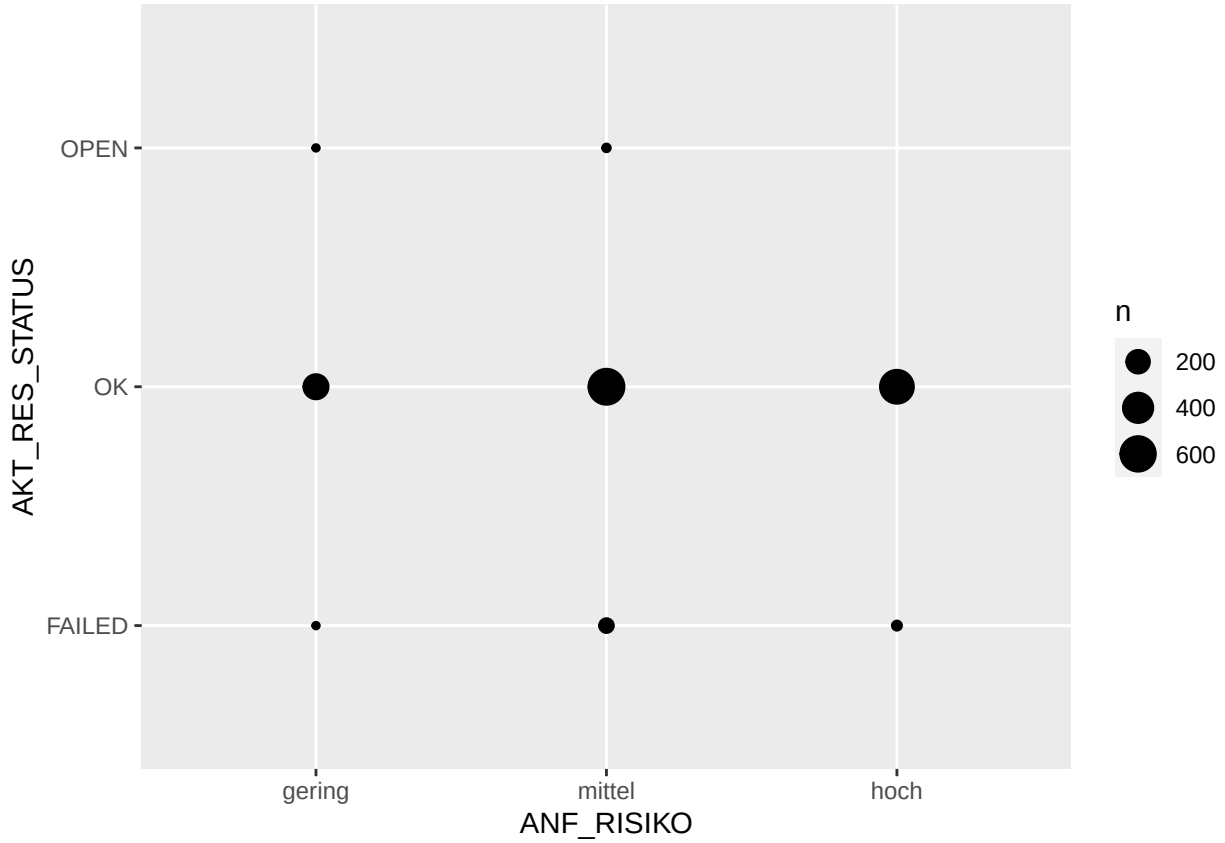
```
# absolute Haeufigkeiten von RES_STATUS nach RISIKO
par(mfrow=c(1,3))
ggplot(data = ds1) + geom_count(mapping = aes(x = ANF_RISIKO, y = AKT_RES_STATUS))
```




```
ggplot(data = ds2) + geom_count(mapping = aes(x = ANF_RISIKO, y = AKT_RES_STATUS))
```



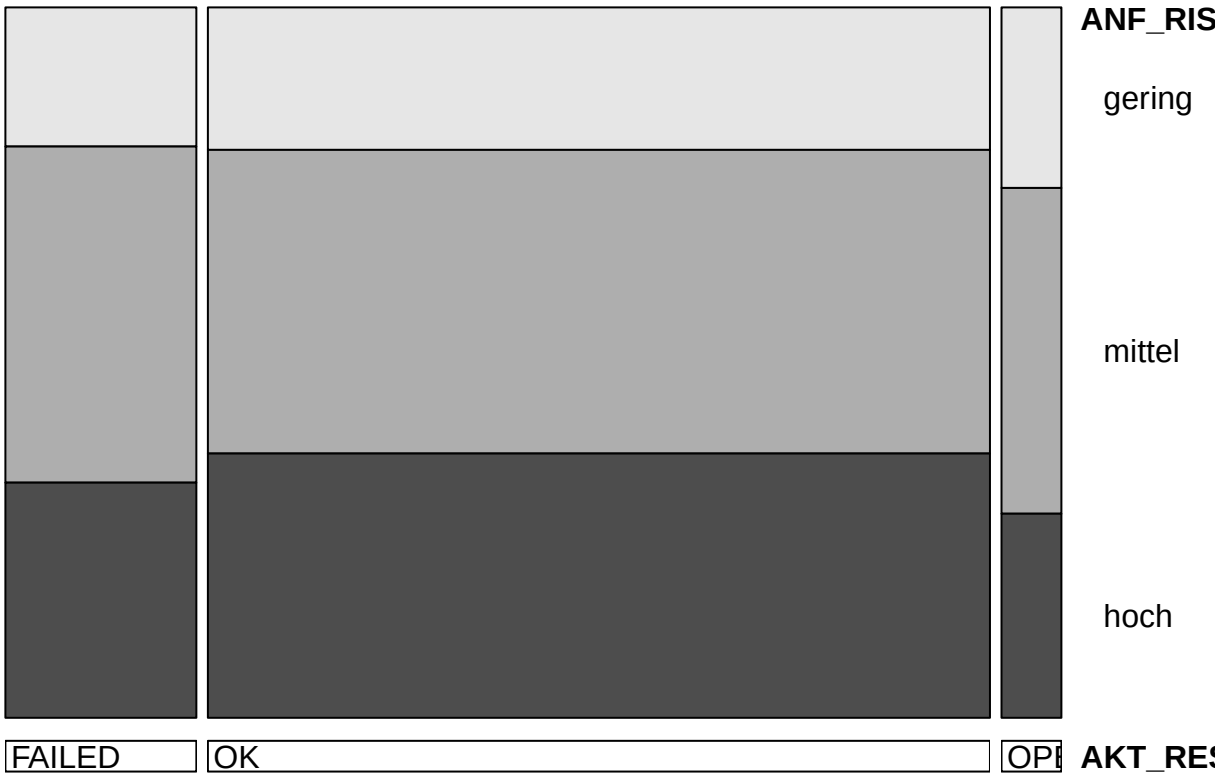

```
ggplot(data = ds3) + geom_count(mapping = aes(x = ANF_RISIKO, y = AKT_RES_STATUS))
```




```
par(mfrow=c(1,1))
```

```
# relative Anteile von RES_STATUS nach RISIKO
par(mfrow=c(1,3))
doubledecker(ANF_RISIKO ~ AKT_RES_STATUS, data = ds1)
```

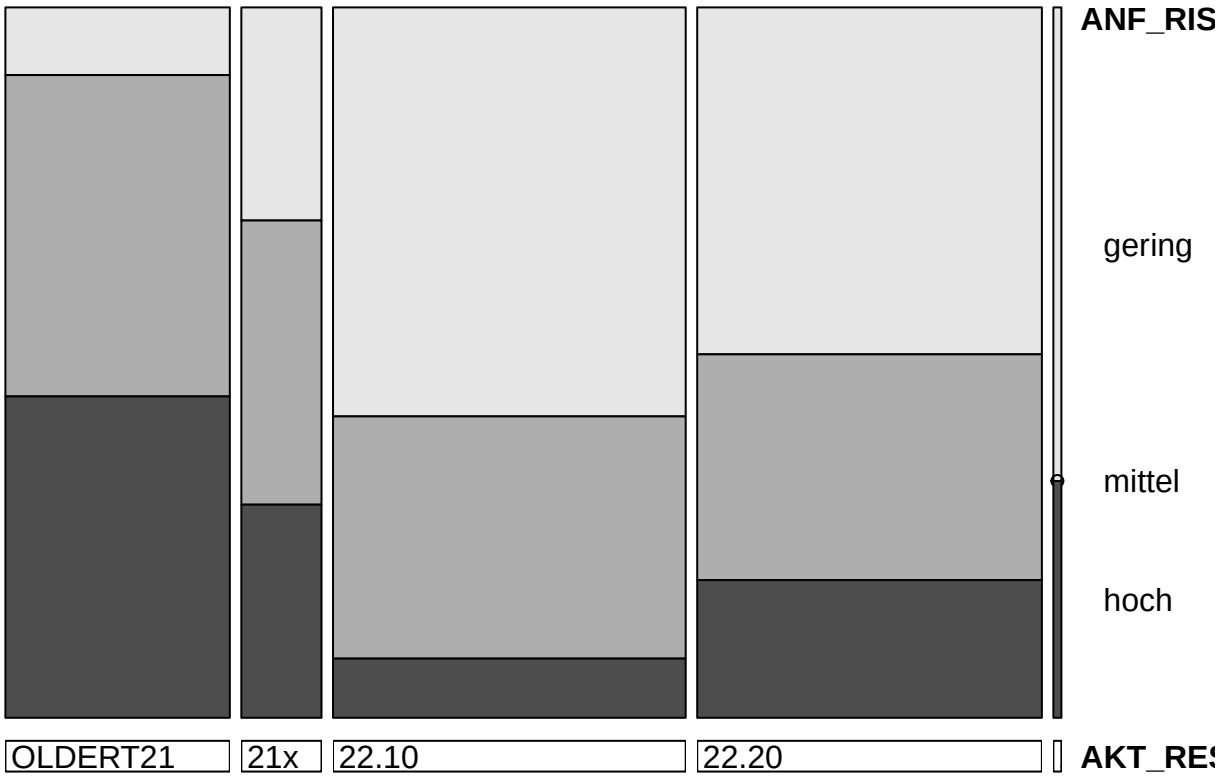

```
doubledecker(ANF_RISIKO ~ AKT_RES_STATUS, data = ds2)
```



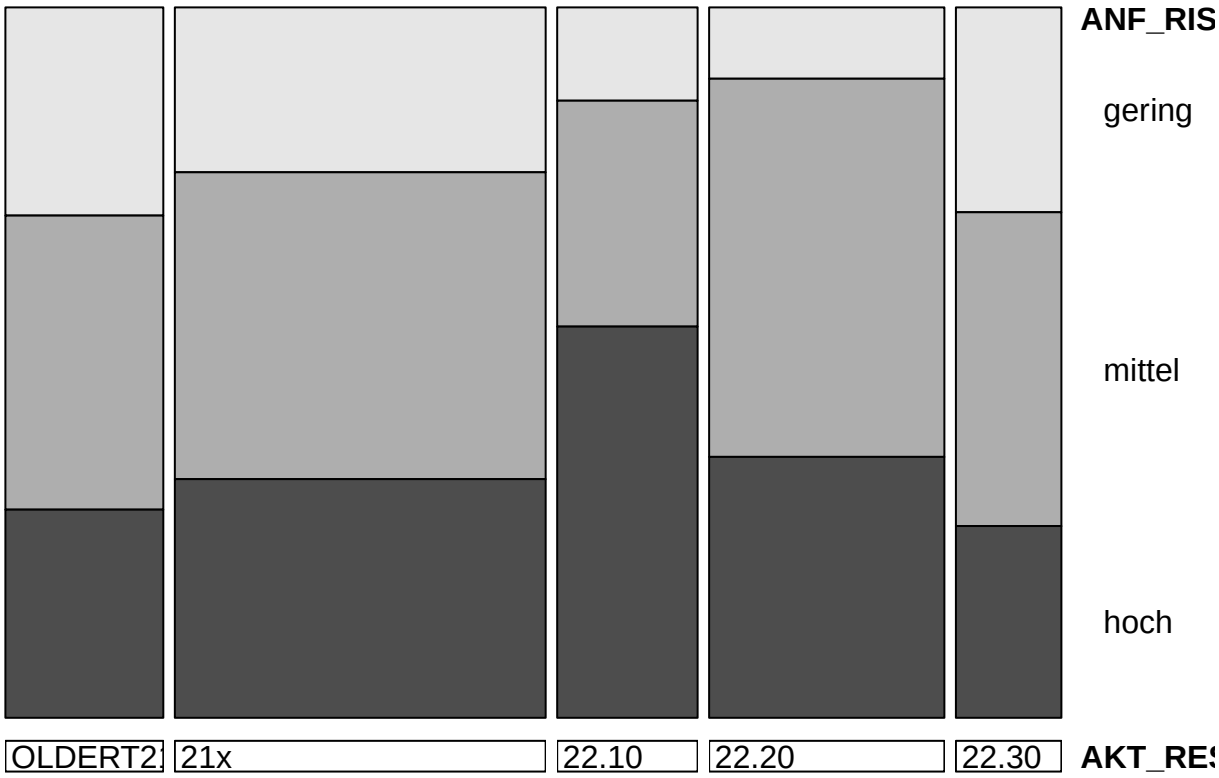

```
doubledecker(ANF_RISIKO ~ AKT_RES_STATUS, data = ds3)
```


```
par(mfrow=c(1,1))
```

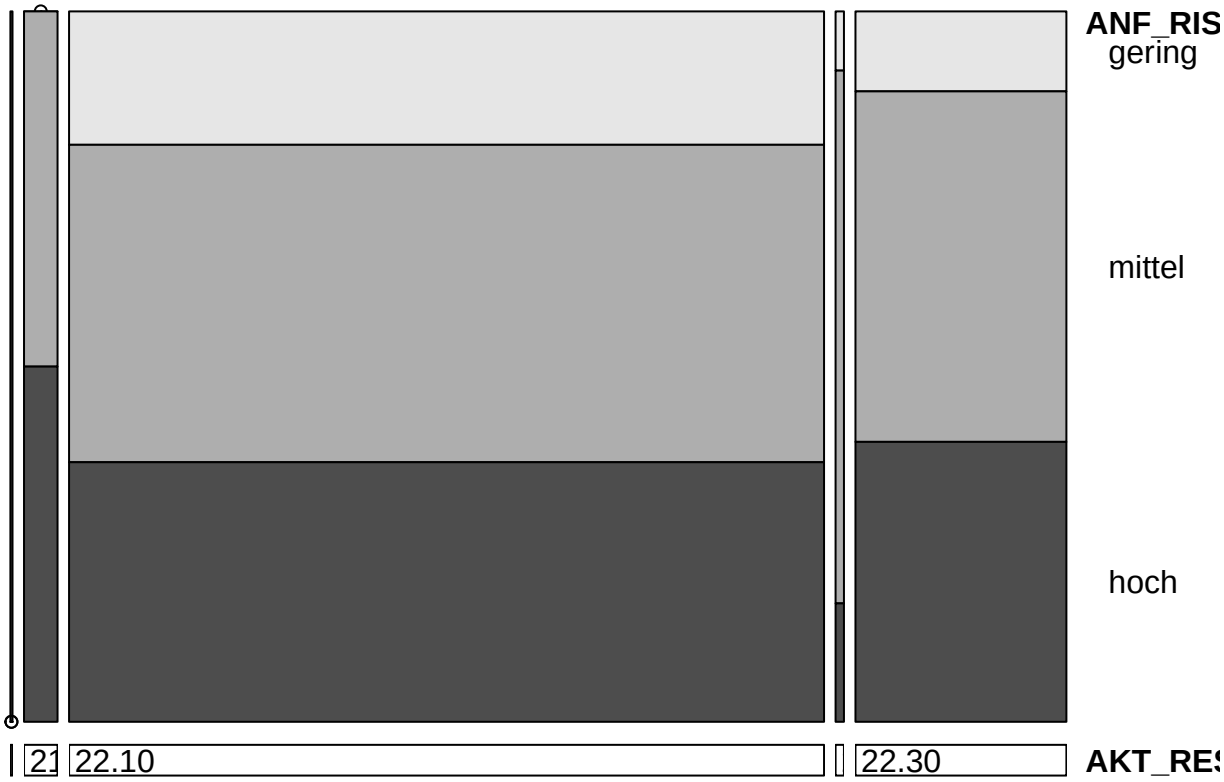
```
# relative Anteile von RES_RELEASE nach RISIKO
par(mfrow=c(1,3))
doubledecker(ANF_RISIKO ~ AKT_RES_RELEASE, data = ds1)
```




```
doubledecker(ANF_RISIKO ~ AKT_RES_RELEASE, data = ds2)
```




```
doubledecker(ANF_RISIKO ~ AKT_RES_RELEASE, data = ds3)
```



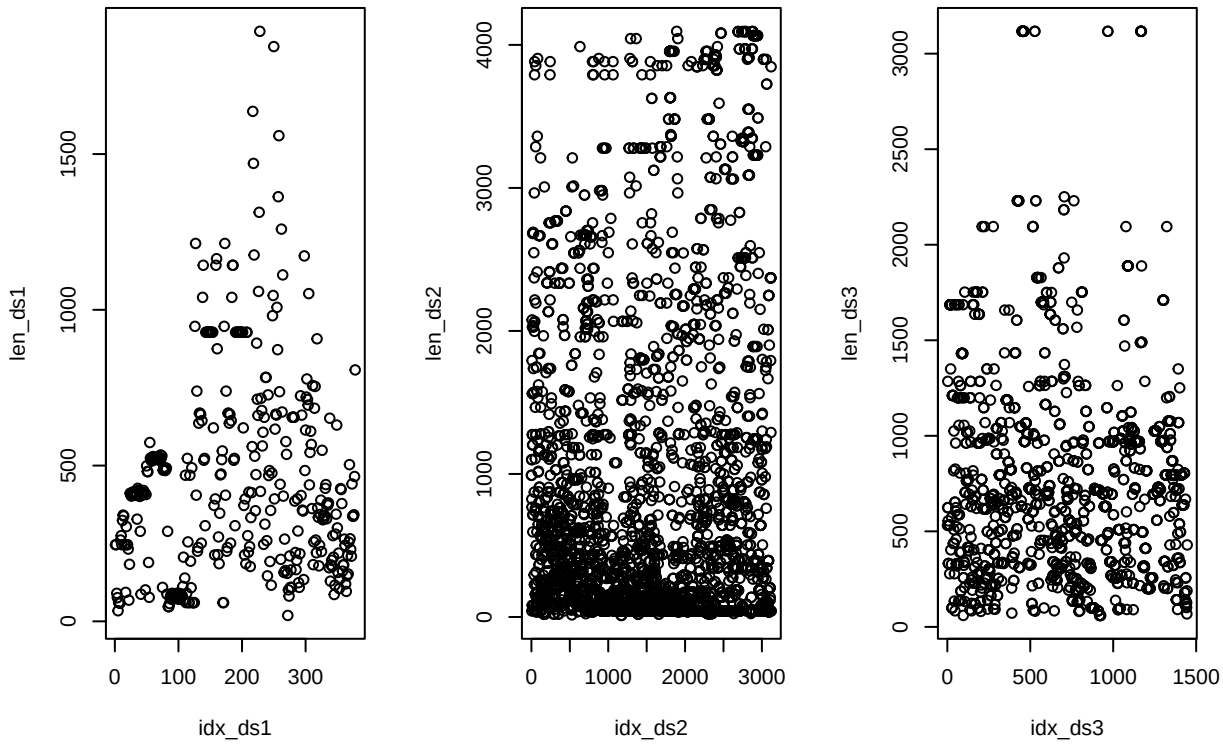

```
par(mfrow=c(1,1))
```

metrische Variable

Da die textuellen Beschreibungen aus der Spalte `ANF_BESCHREIBUNG` zur Klassifikation verwendet werden sollen, wird im folgenden Schritt die Länge der Beschreibungen dargestellt. Im Datenset `ds1` sind die Beschreibungen im Vergleich zu `ds2` und `ds3` sehr kurz. Zu Vergleichszwecken wird aber auch dieses Datenset für die Klassifikation verwendet.

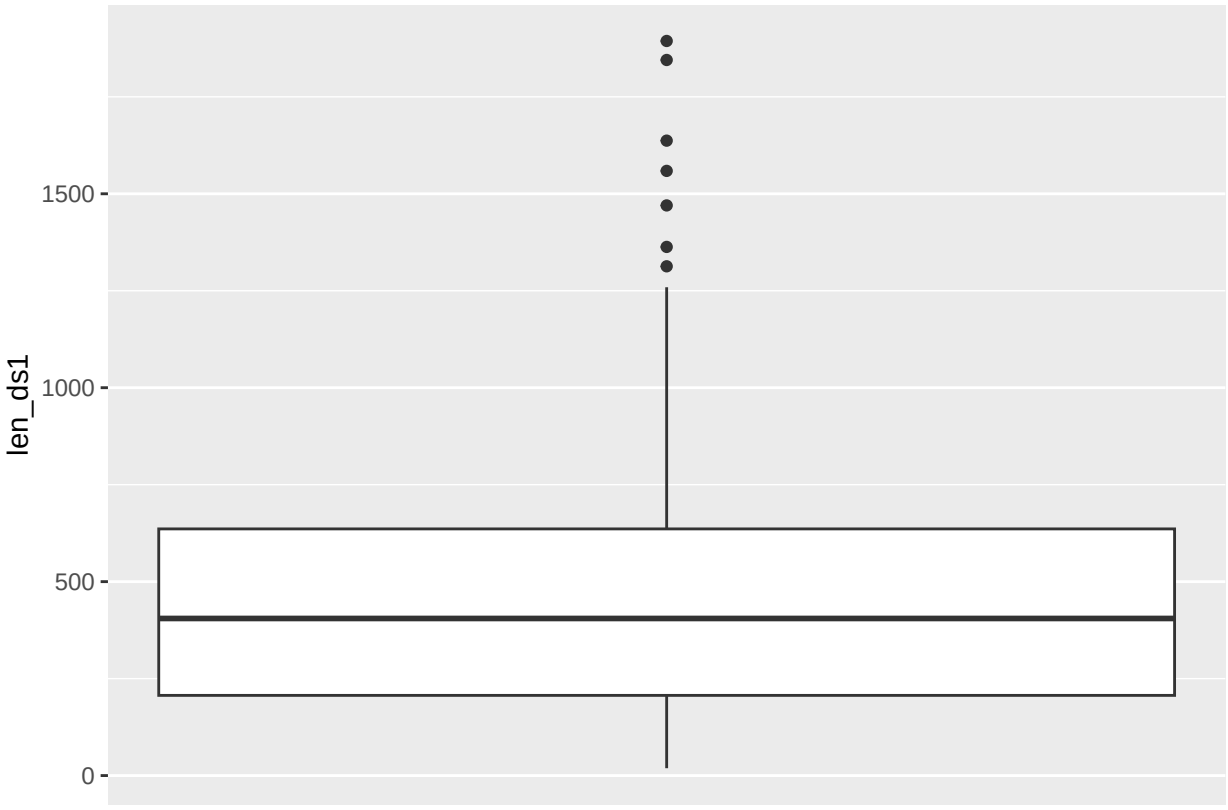
```
# Analyse von Laenge der Beschreibungen
idx_ds1 <- vector()
len_ds1 <- vector()
for (i in 1:length(ds1[, 3])) {
  idx_ds1[i] <- i
  len_ds1[i] <- nchar(ds1[i,3])
}
idx_ds2 <- vector()
len_ds2 <- vector()
for (i in 1:length(ds2[, 3])) {
  idx_ds2[i] <- i
  len_ds2[i] <- nchar(ds2[i,3])
}
idx_ds3 <- vector()
len_ds3 <- vector()
for (i in 1:length(ds3[, 3])) {
  idx_ds3[i] <- i
  len_ds3[i] <- nchar(ds3[i,3])
```

```
}
par(mfrow=c(1,3))
plot(idx_ds1, len_ds1)
plot(idx_ds2, len_ds2)
plot(idx_ds3, len_ds3)
```

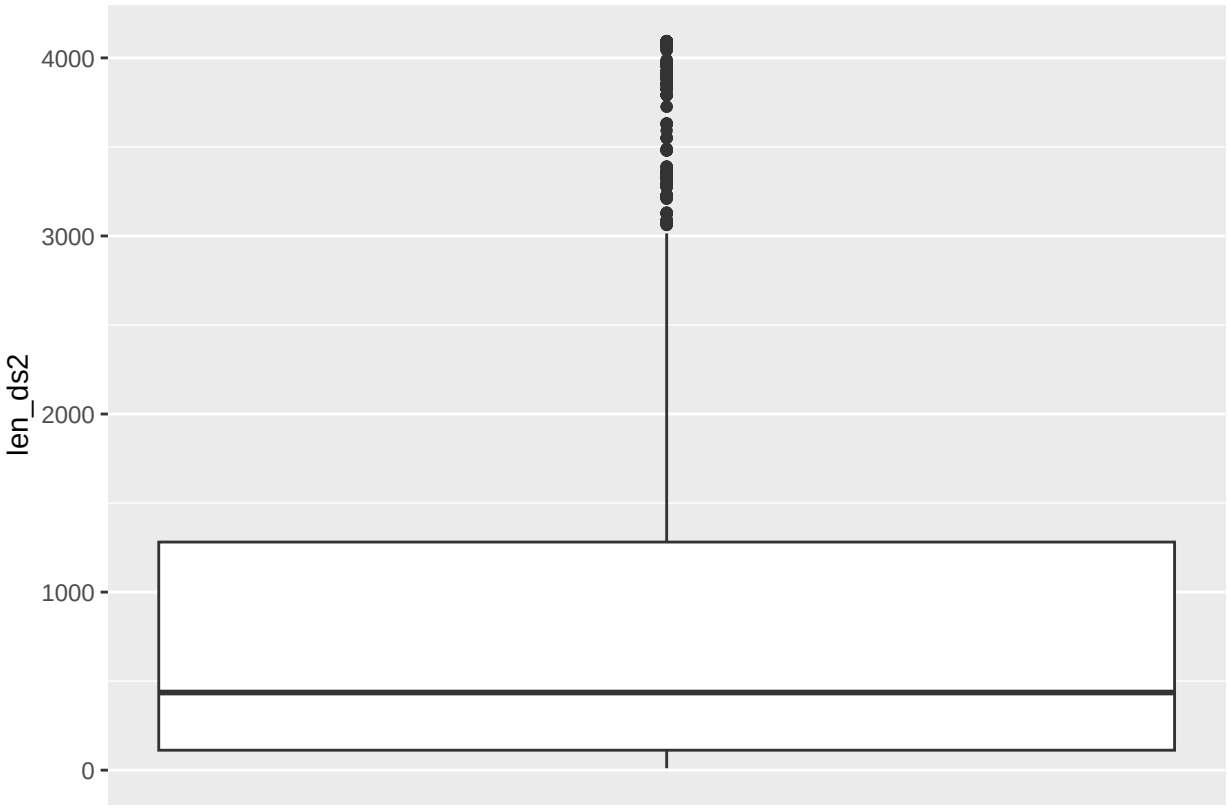



```
par(mfrow=c(1,1))
```

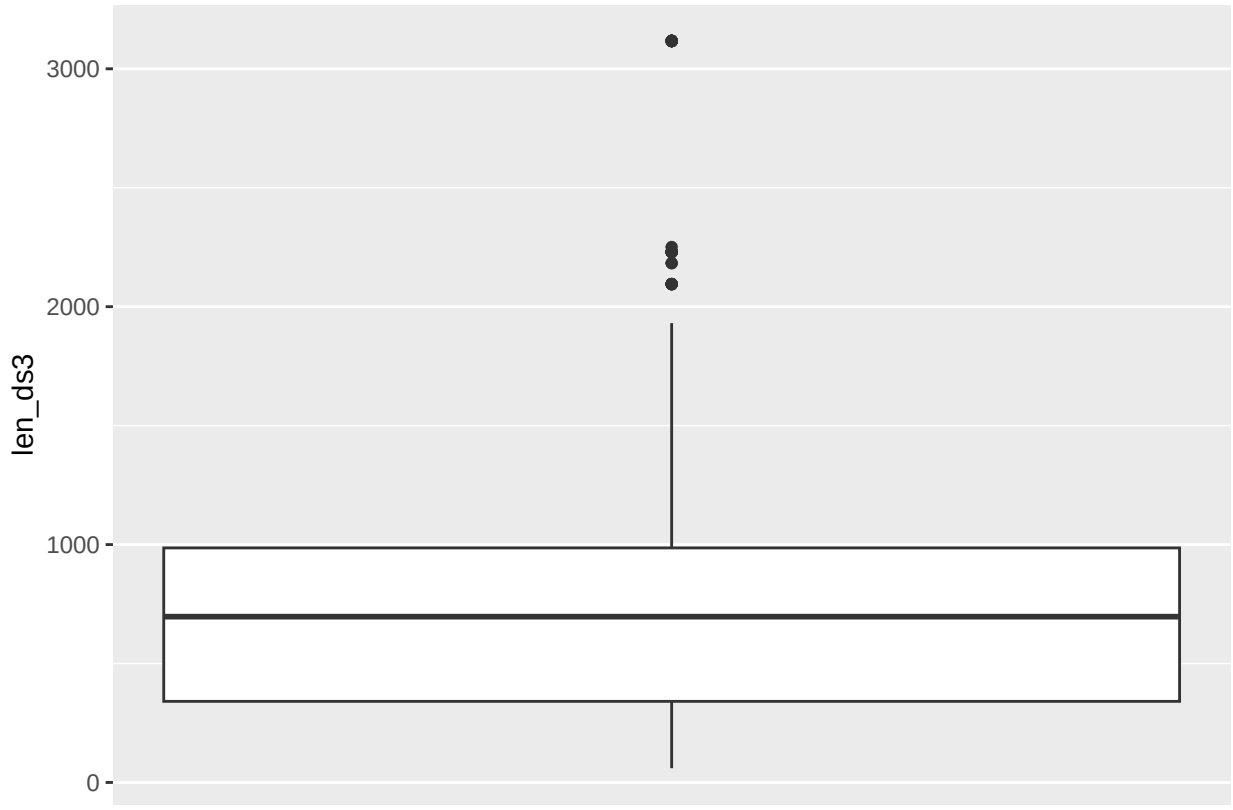
```
# Boxplots der Laenge der Beschreibungen
df1 <- as.data.frame(cbind(idx_ds1, len_ds1))
df2 <- as.data.frame(cbind(idx_ds2, len_ds2))
df3 <- as.data.frame(cbind(idx_ds3, len_ds3))

par(mfrow=c(1,3))
ggplot(df1, aes(1, len_ds1)) +
geom_boxplot() +
xlab("") + scale_x_continuous(breaks = NULL)
```




```
ggplot(df2, aes(1, len_ds2)) +
geom_boxplot() +
xlab("") + scale_x_continuous(breaks = NULL)
```




```
ggplot(df3, aes(1, len_ds3)) +
geom_boxplot() +
xlab("") + scale_x_continuous(breaks = NULL)
```




```
par(mfrow=c(1,1))
```

Zusammenfassung

Abschließend werden die drei Datensets noch einmal zusammengefasst und jeweils ein Überblick ausgegeben.

```
summary(ds1)
```

```
##     ANF_ID             ANF_NAME         ANF_BESCHREIBUNG    ANF_RISIKO
##  Length:378         Length:378         Length:378         gering:158
##  Class :character   Class :character   Class :character   mittel:136
##  Mode  :character   Mode  :character   Mode  :character   hoch  : 84
##
##
##
##   TF_ABDECKUNG    AKT_RES_STATUS AKT_RES_RELEASE
##  Min.   :  0.00   FAILED: 12     OLDERT21: 84
##  1st Qu.: 50.00   OK    :359     21x     : 30
##  Median :100.00   OPEN  :  7     22.10   :132
##  Mean   : 80.56                  22.20   :129
##  3rd Qu.:100.00                  22.30   :  3
##  Max.   :100.00
```

```
summary(ds2)
```

```
##     ANF_ID             ANF_NAME         ANF_BESCHREIBUNG
##  Length:3121        Length:3121        Length:3121
```

```
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  ANF_FEHLERWAHRSCHEINLICHKEIT ANF_FEHLERKOSTEN  ANF_RISIKO    TF_ABDECKUNG
##  gering: 998                  gering: 705      gering: 633   Min.   : -0.70
##  mittel:1307                  mittel: 856      mittel:1366   1st Qu.:  3.45
##  hoch  : 816                  hoch  :1560      hoch  :1122   Median : 16.65
##                                                              Mean   : 29.79
##                                                              3rd Qu.: 50.00
##                                                              Max.   :100.00
##  AKT_RES_STATUS AKT_RES_RELEASE
##  FAILED: 577    OLDERT21: 488
##  OK    :2363    21x     :1146
##  OPEN  : 181    22.10   : 434
##                 22.20   : 727
##                 22.30   : 326
##
```

```
summary(ds3)
```

```
##     ANF_ID             ANF_NAME         ANF_BESCHREIBUNG
##  Length:1446        Length:1446        Length:1446
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  ANF_FEHLERWAHRSCHEINLICHKEIT ANF_FEHLERKOSTEN  ANF_RISIKO   TF_ABDECKUNG
##  gering:233                   gering:178       gering:241   Min.   :  0.00
##  mittel:966                   mittel:804       mittel:665   1st Qu.:  6.67
##  hoch  :247                   hoch  :464       hoch  :540   Median : 14.29
##                                                             Mean   : 23.20
##                                                             3rd Qu.: 33.33
##                                                             Max.   :100.00
##  AKT_RES_STATUS AKT_RES_RELEASE
##  FAILED:  48    OLDERT21:   3
##  OK    :1395    21x     :  48
##  OPEN  :   3    22.10   :1081
##                 22.20   :  12
##                 22.30   : 302
##
```