vincent_grunert_mietspiegel.R

vincent

2021-11-06

```
library(readxl)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
options(scipen = 10)
path <- "/home/vincent/Dropbox/University/Statistik/KURSE/CURRENT/master/stat_fallstudien/Beispiel_Miets

miete <- read_xlsx(path, sheet = 2)



# Ich bin Immobilienmakler und betreibe in Wien ein Firma. Wir wollen eine genauere Analyse des Marktes

# Uns interessiert, was sind die Key Treiber für den Mietpreis bei privater Wohnungsmiete in unserer St

# Ich beauftrage Sie ein entsprechendes Untersuchungsdesign zu erstellen.

# Univariate Analysesn
# a) Kennzahlen
#     Mittelwert, Standardabweichung, Schiefe, Minimum, 1.Quantil, Median, 3.Quantil, Maximum
#     Modus, eventuell 2., 3. etc. größte Werte
# b) Grafische Darstellung
# c) Datenkontrolle, Qualitätskontrolle
# d) Ziel: machen Sie sich ein Bild von den Objekten, die
# Sie vor sich haben.


attach(miete)

names(miete)
```

```
##  [1] "nm"       "wfl"      "rooms"    "bj"       "wohngut"  "wohnbest"
##  [7] "ww0"      "zh0"      "badkach0" "badextra" "kueche"
```

```r
# erstelle eine neue variable: alter = 2003 - bj

# daten kontrollieren
# missing values?
any(is.na(miete))
```

```
## [1] FALSE
```

```r
# keine zahl?
any(apply(miete, 2, is.nan))
```

```
## [1] FALSE
```

```r
# sind alles zahlen?
all(apply(miete, 2, is.numeric))
```

```
## [1] TRUE
```

```r
# dh wir haben auch keine string werte in den daten

# negarive zahlen (duerften nicht vorkommen)
any(miete < 0)
```

```
## [1] FALSE
```

```r
# wir wissen bereits dass keine Zahl kleiner wie null ist, keine missing values und auch keine undendli
# nm und wohnflaeche sind somit zulaessig
# rooms kann nur naruerliche zahlen annehmen
names(table(rooms))
```

```
## [1] "1" "2" "3" "4" "5" "6"
```

```r
# ok das ist auch in ordnung

# baujahr kann verschiedene werte annehmen die sinn machen, daher muss
# hier eine idividuelle beurteilung erfolgen
names(table(bj))
```

```
##  [1] "1918"   "1924"   "1939"   "1948"   "1957"   "1957.5" "1960"   "1966"
##  [9] "1967"   "1968"   "1969"   "1970"   "1971"   "1972"   "1973"   "1974"
## [17] "1975"   "1976"   "1977"   "1978"   "1979"   "1980"   "1981"   "1982"
## [25] "1983"   "1984"   "1985"   "1986"   "1987"   "1988"   "1989"   "1990"
## [33] "1991"   "1992"   "1993"   "1994"   "1995"   "1996"   "1997"   "1998"
## [41] "1998.5" "1999"   "2000"   "2001"
```

```r
# wir sehen dass es sich bis auf zwei beobachtungen um ganzzahlig werte die zwischen 1918 - 2001 liegen

# alle anderen variablen haben nur null und eins als definitionsbereich. dies laesst sich leicht testen
is_binary <- function(x) x == 0 | x == 1
all(apply(miete[,5:ncol(miete)], 2, function(x) all(is_binary(x))))
```

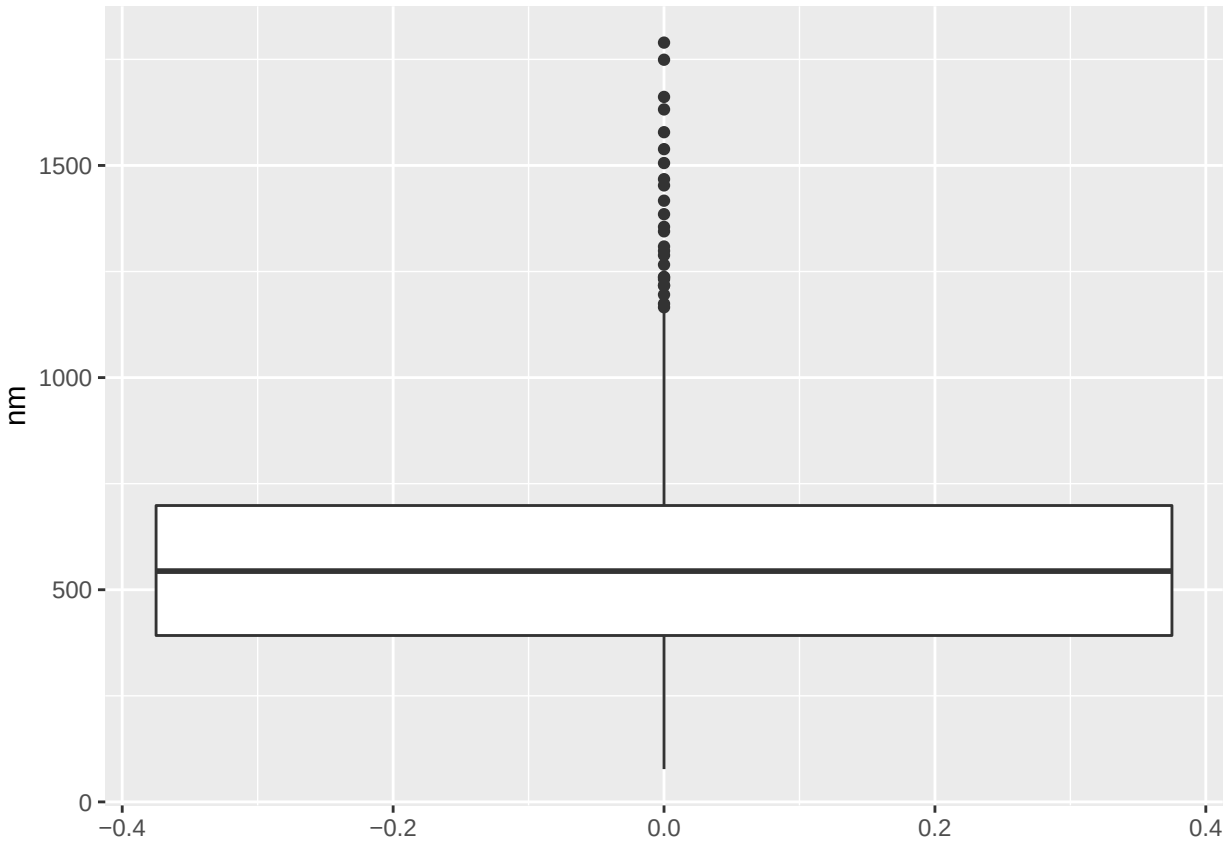
```
## [1] TRUE
```

```r
# perfekt, alles binaries

# 1.5 * iqr regel
get_outlier <- function(x) {
    iqr <- quantile(x, 0.75) - quantile(x, 0.25)
    which( x > quantile(x, 0.75) + 1.5 * iqr | x < quantile(x, 0.25) - 1.5 * iqr)
```

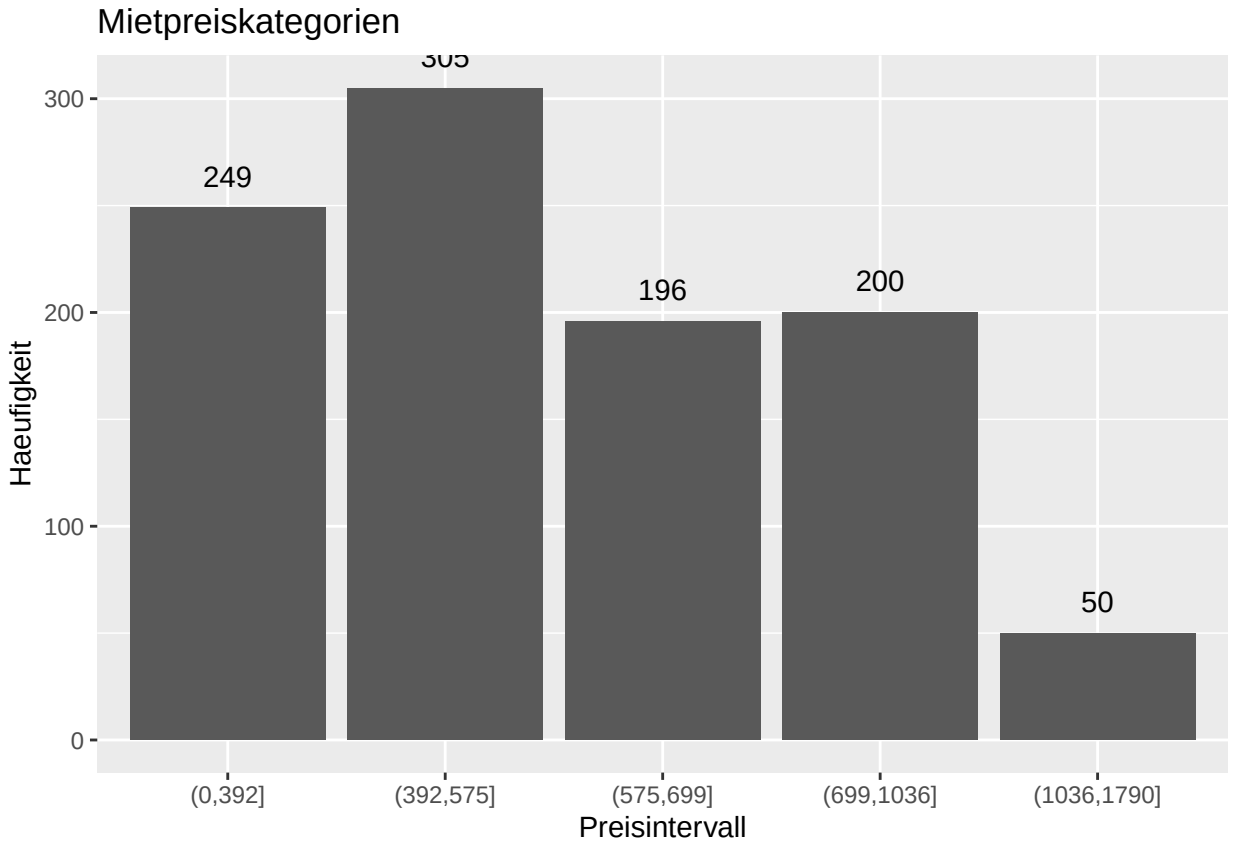
```r
}

has_outlier <- function(x){
    if(length(get_outlier(x)) > 0) return(TRUE)
    else{return(FALSE)}
}


# somit koennen wir festhalten, dass die daten ihren zulaessigen werten entsprechen
# und auch keine weitere bereinigung/entfernung der daten notwendig ist.

alter <- 2003 - bj

# univariate analysen

# nettomiete
# "grobstistiken"
summary(nm)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   77.31  392.15  543.93  574.87  698.52 1789.55
```

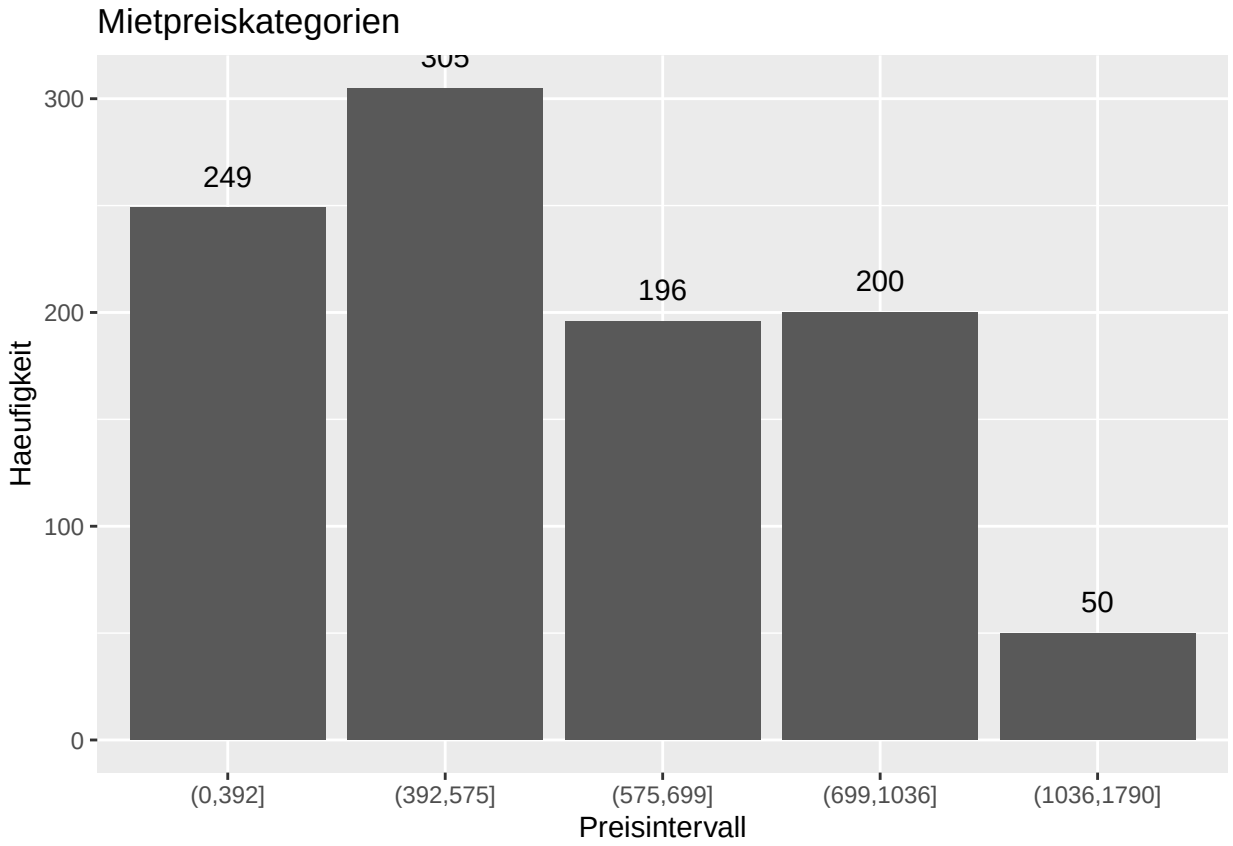
```r
# wir gruppieren die daten in verschiedene kategorien um die unterschiede zu verdeutlichen
# cat_nm <- cut(nm, c(min(nm), quantile(nm, 0.25), mean(nm), quantile(nm, 0.75), quantile(nm, 0.95), ma

# zahlen gerundet
nm_stats <- c(0, round(quantile(nm, 0.25)), round(mean(nm)), round(quantile(nm, 0.75)), round(quantile(n

cat_nm <- cut(nm, nm_stats, dig.lab = 5)

# anzahl an beobachtungen in den jeweiligen kategorien
table_nm_cat <- table(cat_nm)

# speicher im df fuer die grafische aufbearbeitung
df_nm <- data.frame(table_nm_cat)

# cumsums
cumsum_cat_nm <- cumsum(table_nm_cat)
names(cumsum_cat_nm) <- paste("Nmiete <=", nm_stats[-1])
df_cumsum_nm <- data.frame(Cat = names(cumsum_cat_nm), Freq = cumsum_cat_nm)


# boxplot: einige ausreisser nach oben
ggplot(miete) + geom_boxplot(aes(y = nm))
```



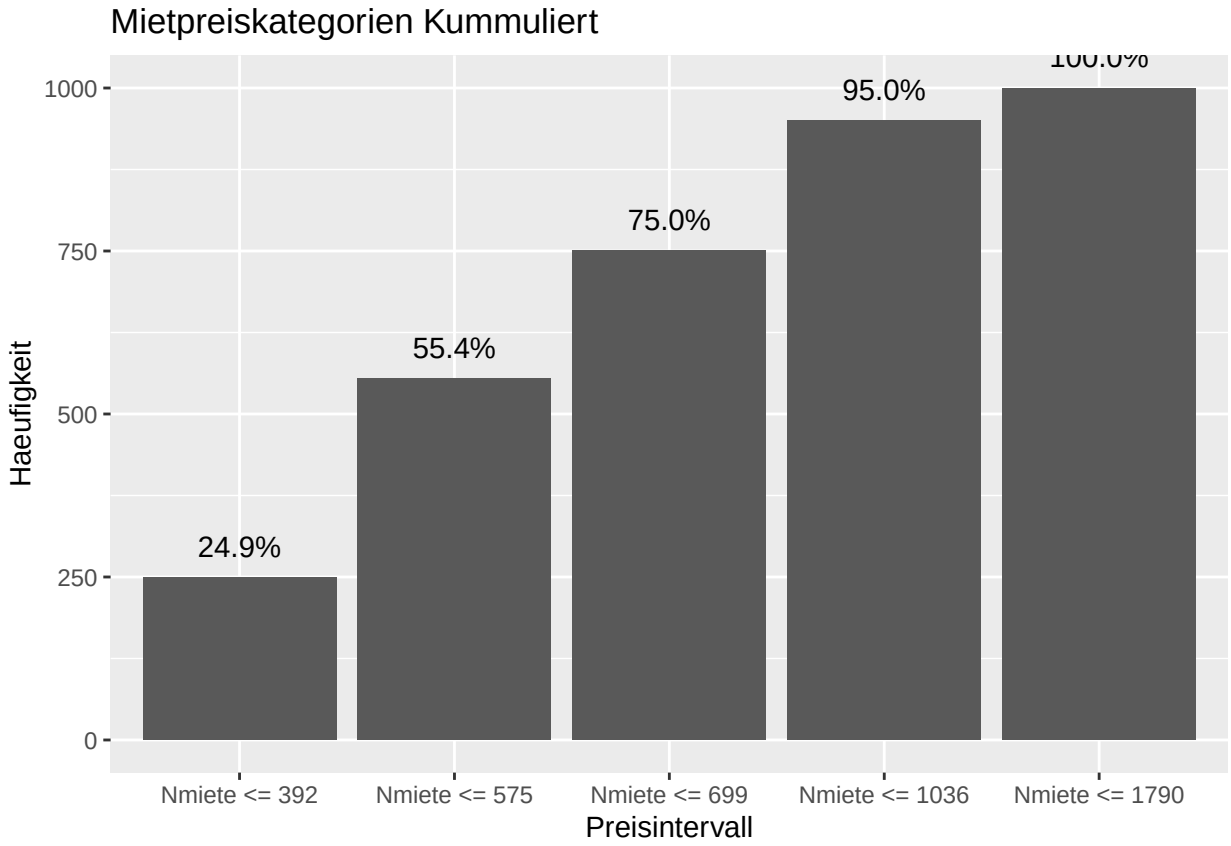

```
# histogram
ggplot(df_nm, aes(x=cat_nm, y = Freq)) + geom_bar(stat = "identity") + geom_text(aes(label = Freq), vju
       x ="Preisintervall", y = "Haeufigkeit")
```




```
# h <- hist(nm)
# histo <- data.frame(Intervals = cut(h$breaks, breaks = h$breaks, dig.lab = 5)[-1], Counts = h$counts)

# ggplot(histo, aes(x = Intervals, y = Counts)) + geom_bar(stat = "identity")
# ggplot(miete) + geom_histogram(aes(x = nm), bins = 10)

# histogram der kategorien
ggplot(df_nm, aes(x=cat_nm, y = Freq)) + geom_bar(stat = "identity") + geom_text(aes(label = Freq), vjus
       x ="Preisintervall", y = "Haeufigkeit")
```




```
# kummulierte haeufigkeiten
df_cumsum_nm %>% arrange(Freq) %>% mutate(name = factor(Cat, levels=names(cumsum_cat_nm))) %>% ggplot(a
```




```
# nicht normalverteilt (abwarten bis wir auf die zeit bedingen)
shapiro.test(nm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  nm
## W = 0.93783, p-value < 2.2e-16
```

```
# ausreisser
if(has_outlier(nm)) cbind(index = get_outlier(nm), value = nm[get_outlier(nm)])
```

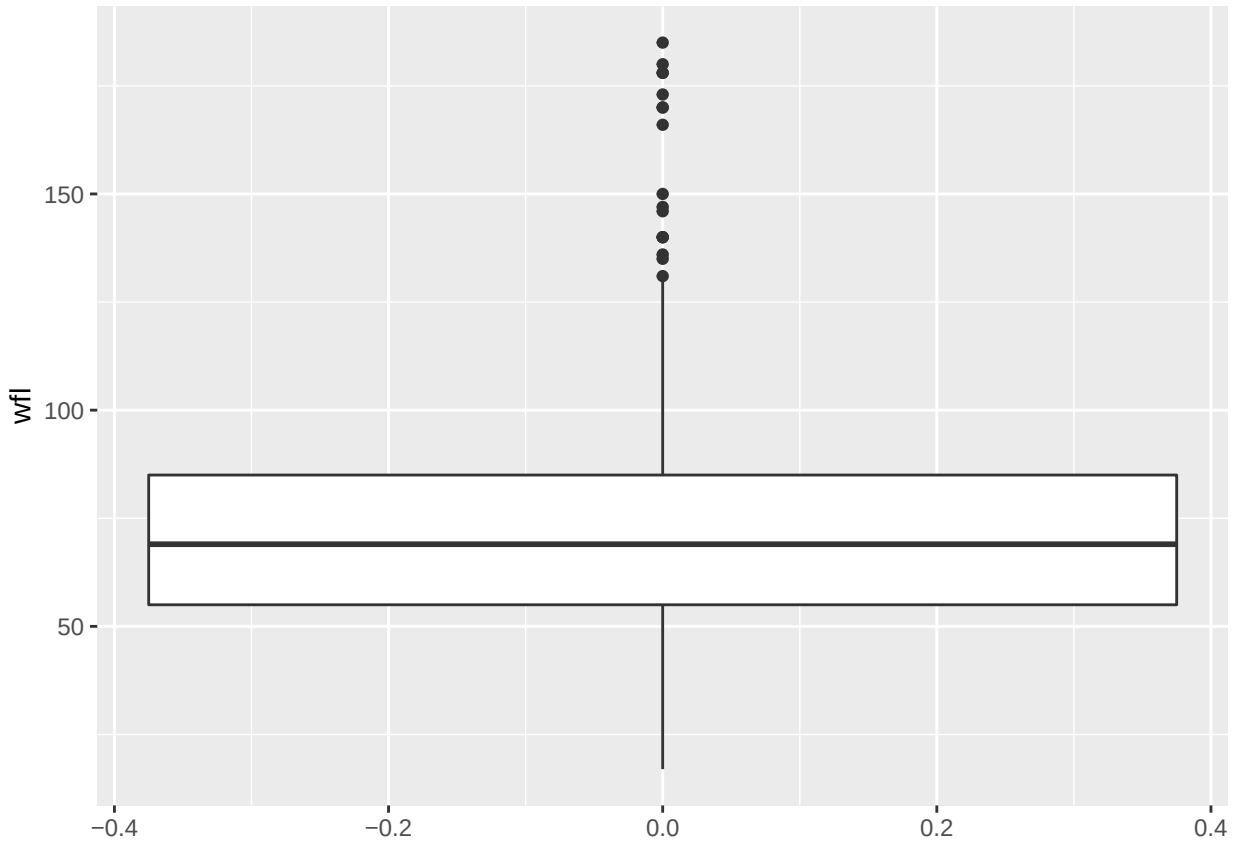
```
##       index   value
##  [1,]    50 1355.28
##  [2,]   114 1749.15
##  [3,]   122 1217.15
##  [4,]   230 1288.48
##  [5,]   267 1195.52
##  [6,]   301 1236.38
##  [7,]   357 1344.72
##  [8,]   371 1789.55
##  [9,]   380 1216.99
## [10,]   436 1165.75
## [11,]   542 1308.94
## [12,]   556 1173.94
## [13,]   667 1298.70
## [14,]   669 1232.00
## [15,]   676 1578.39
```

```
## [16,]   702 1266.05
## [17,]   718 1661.55
## [18,]   722 1452.93
## [19,]   771 1385.12
## [20,]   851 1538.43
## [21,]   864 1467.69
## [22,]   909 1505.66
## [23,]   960 1237.35
## [24,]   962 1416.96
## [25,]   970 1632.03
```

```r
# wohnflaeche
summary(wfl)
```

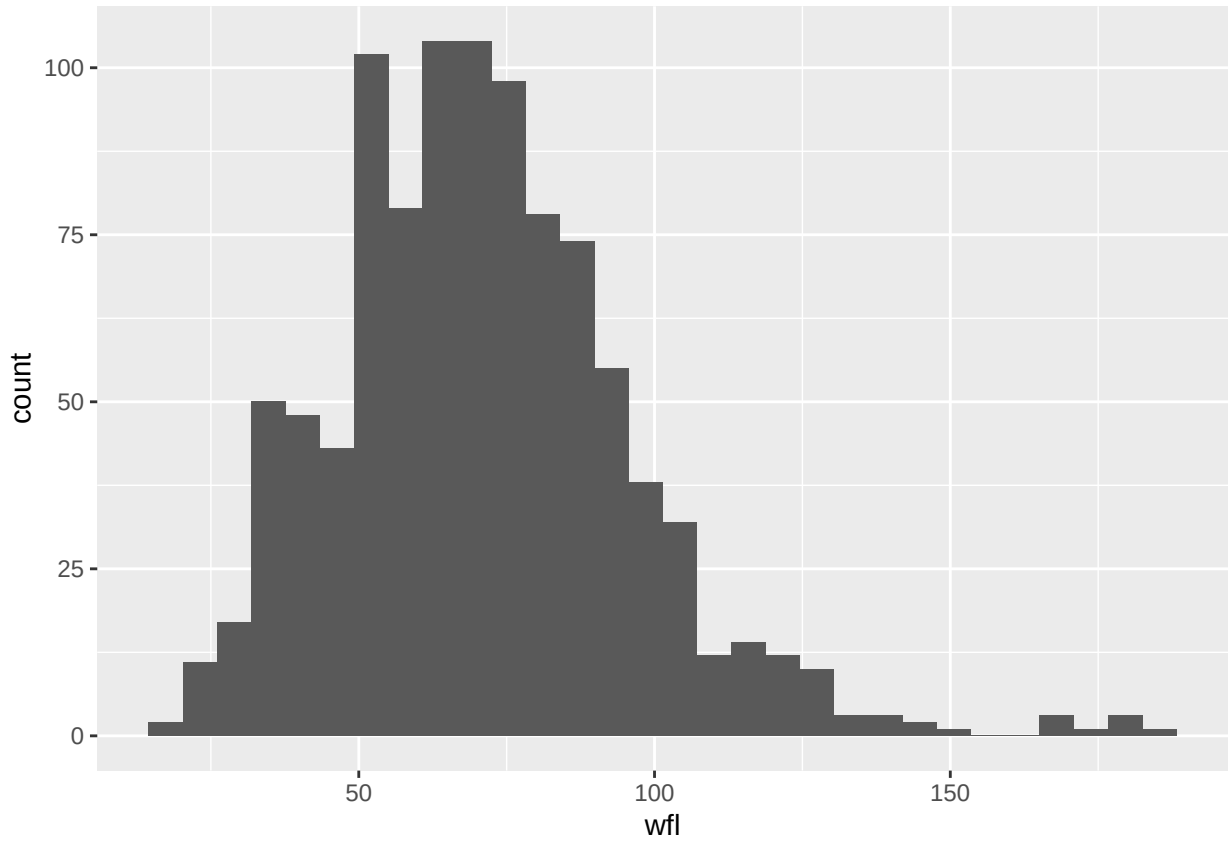
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   55.00   69.00   70.91   85.00  185.00
```

```r
wfl_stats <- c(0, round(quantile(wfl, 0.25)), round(mean(wfl)), round(quantile(wfl, 0.75)), round(quant

cat_wfl <- cut(wfl, wfl_stats)



# haeufigkeiten der kategorien
table_wfl_cat <- table(cat_wfl)

# cusum
cumsum_cat_wfl <- cumsum(table_wfl_cat)
names(cumsum_cat_wfl) <- paste("Wflaeche <=", wfl_stats[-1])
df_cumsum_wfl <- data.frame(Cat = names(cumsum_cat_wfl), Freq = cumsum_cat_wfl)

# speicher in einem data frame fuer die visualisierung
df_wfl <- data.frame(table_wfl_cat)

# einige ausreisser nach oben
ggplot(miete) + geom_boxplot(aes(y = wfl))
```

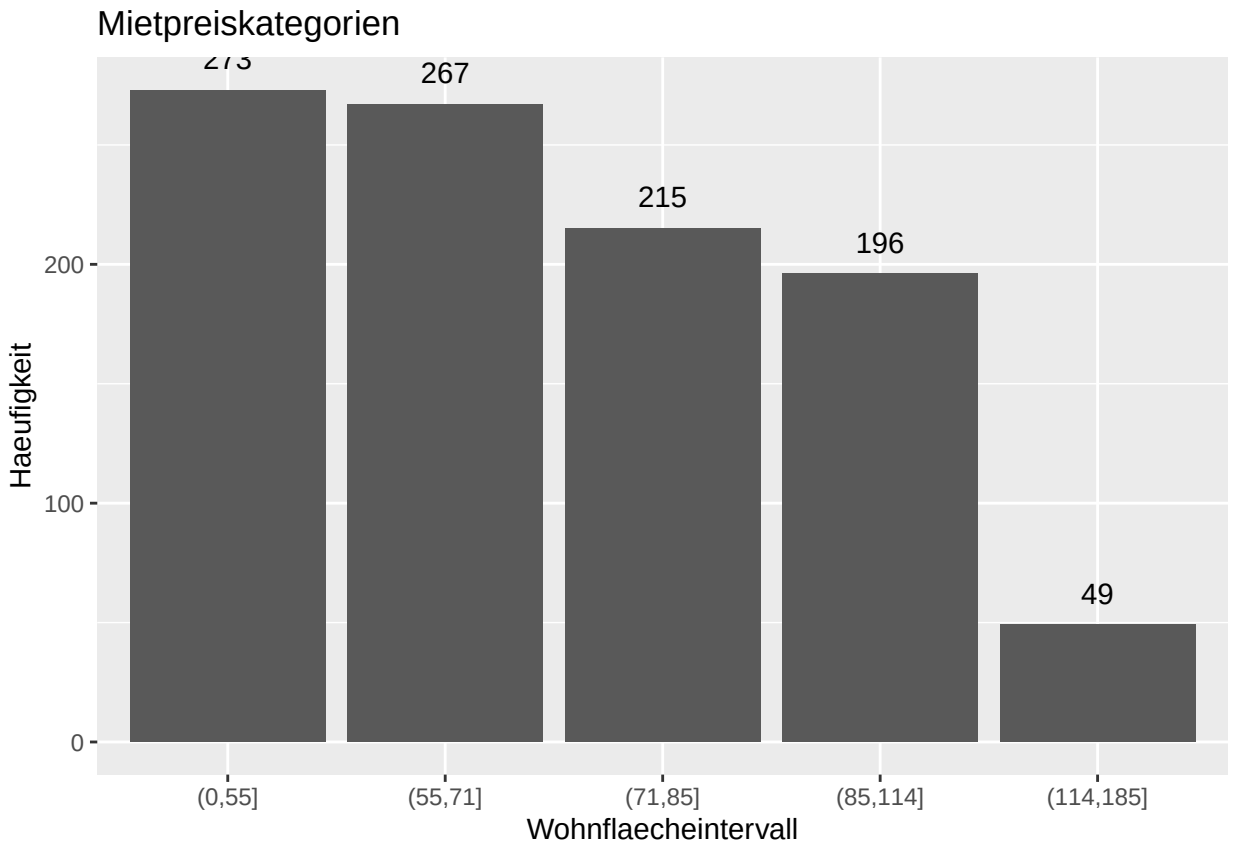



```
# boxplot(wf)

# histogram
ggplot(miete) + geom_histogram(aes(x = wfl))
```

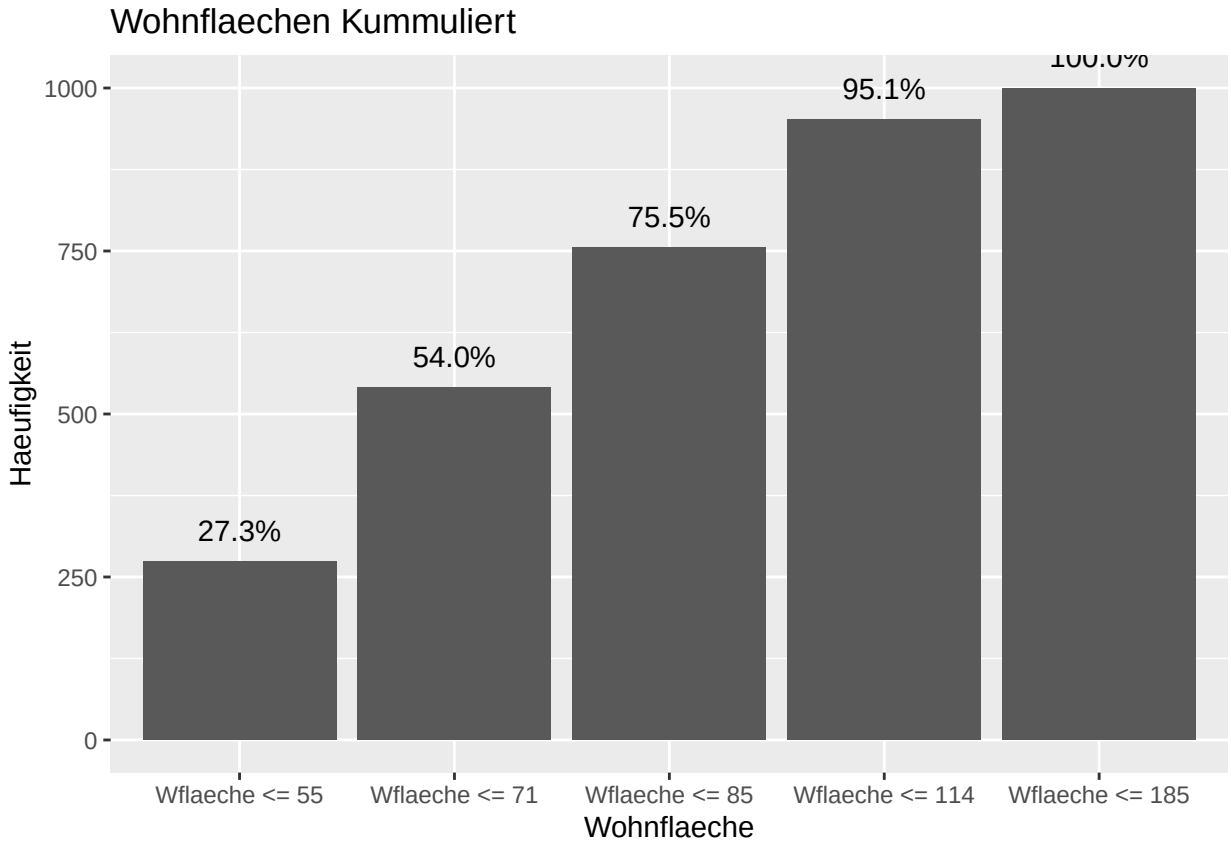
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# histogram der kategorien
ggplot(df_wfl, aes(x=cat_wfl, y = Freq)) + geom_bar(stat = "identity") + geom_text(aes(label = Freq), v
       x ="Wohnflaecheintervall", y = "Haeufigkeit")
```




```
# kummulierte haeufigkeiten
df_cumsum_wfl %>% arrange(Freq) %>% mutate(name = factor(Cat, levels=names(cumsum_cat_wfl))) %>% ggplot
```




```
if(has_outlier(wfl)) cbind(index = get_outlier(wfl), value = nm[get_outlier(wfl)])
```

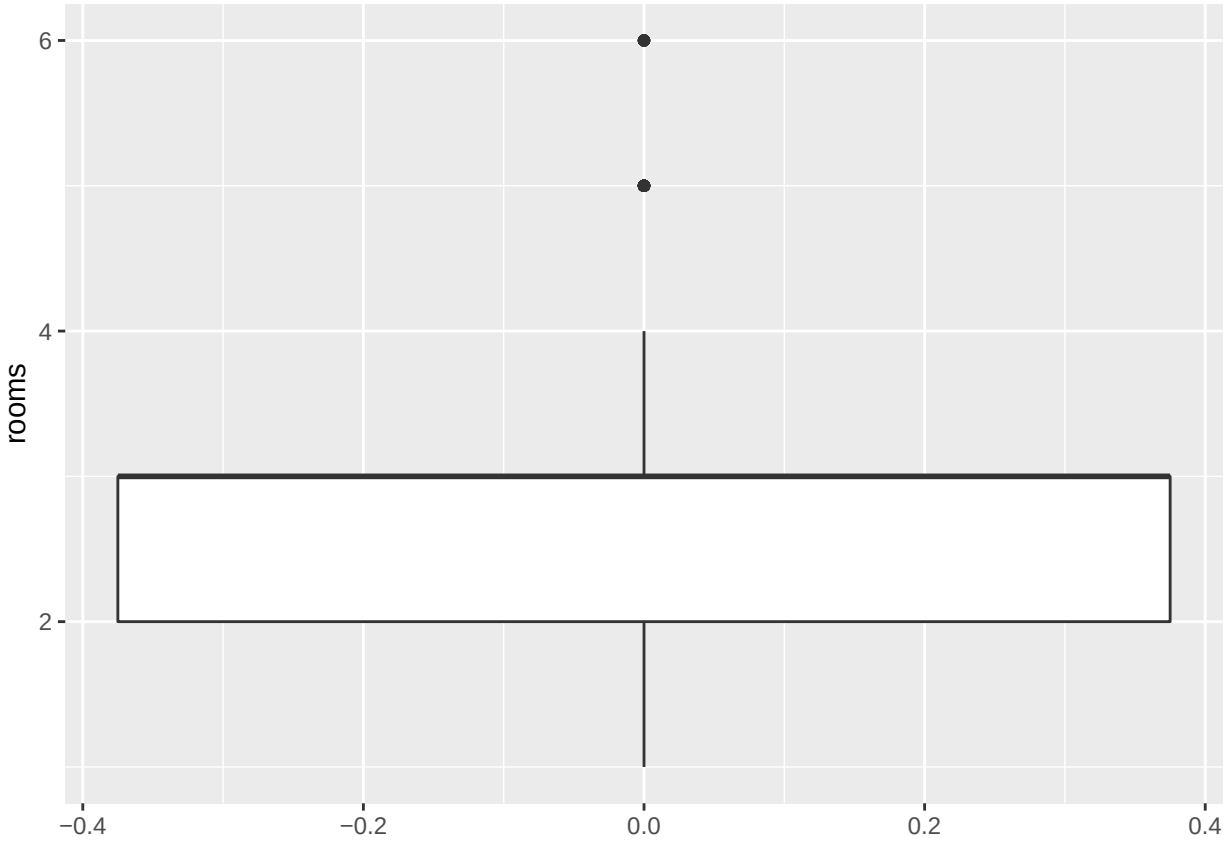
```
##       index   value
##  [1,]    18  796.07
##  [2,]    59 1077.23
##  [3,]    80  526.70
##  [4,]   128  639.13
##  [5,]   203 1113.78
##  [6,]   230 1288.48
##  [7,]   267 1195.52
##  [8,]   543  676.77
##  [9,]   556 1173.94
## [10,]   718 1661.55
## [11,]   851 1538.43
## [12,]   878  821.61
## [13,]   879 1102.00
## [14,]   882  733.77
## [15,]   909 1505.66
## [16,]   932  876.44
## [17,]   962 1416.96
```

```
# zimmer
summary(rooms)
```

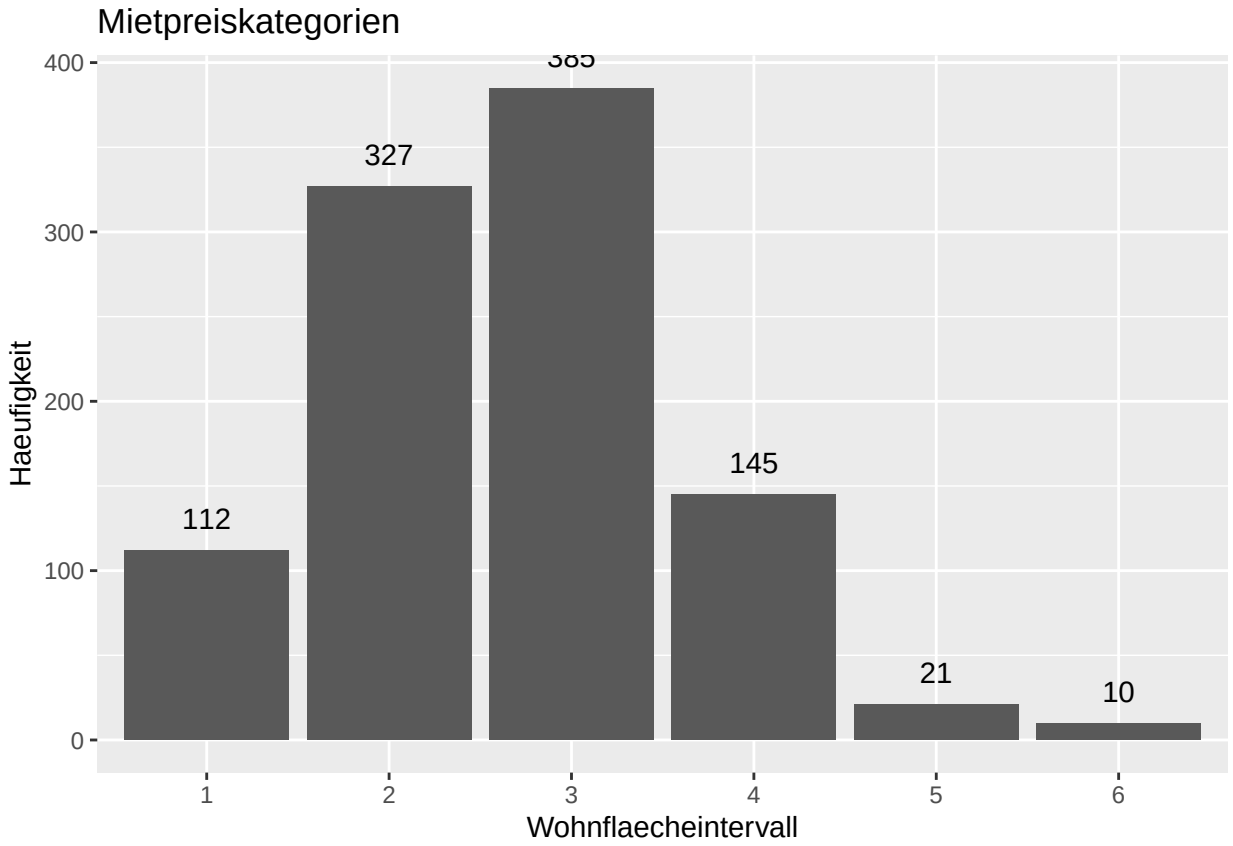
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.666   3.000   6.000
```

```
# haeufigkeiten der kategorien
table_rooms <- table(rooms)
```

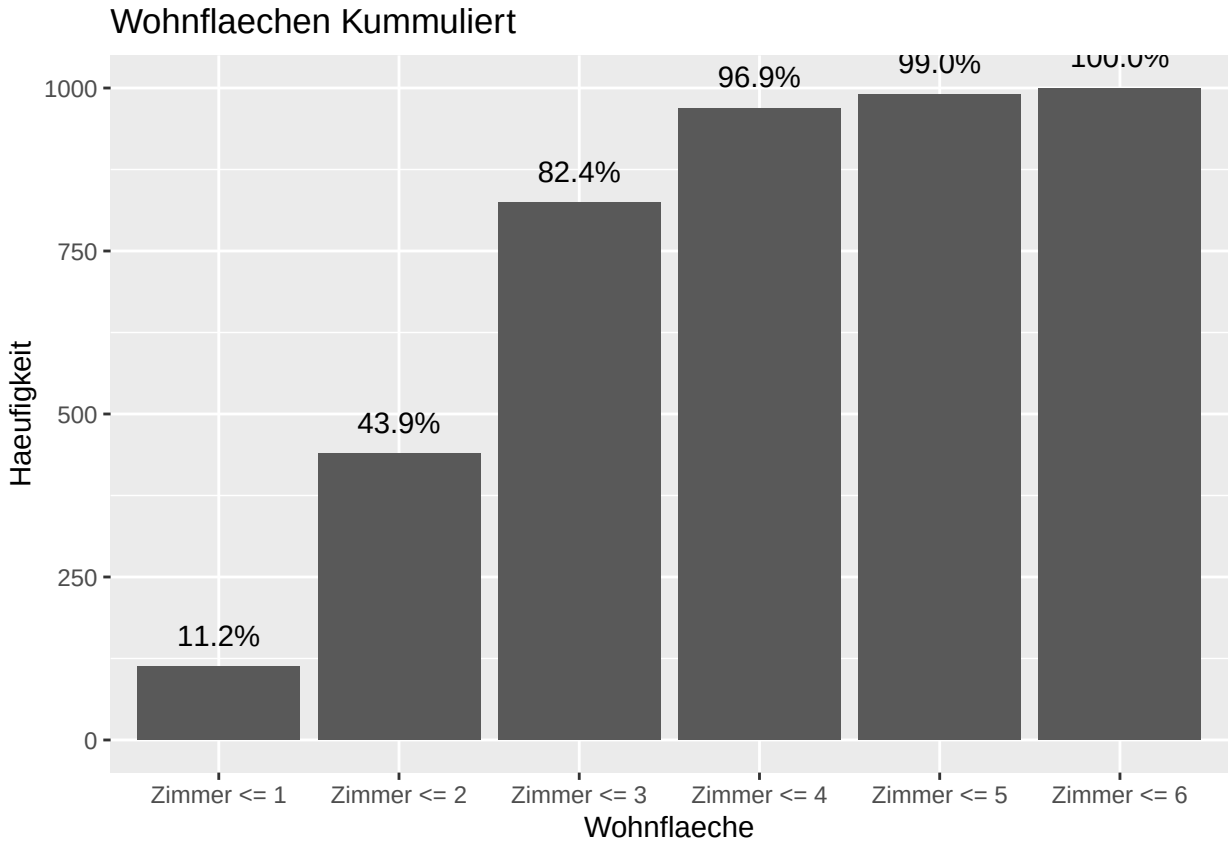
```r
# cusum
cumsum_rooms <- cumsum(table_rooms)
names(cumsum_rooms) <- paste("Zimmer <=", names(table_rooms))
df_cumsum_rooms <- data.frame(Cat = names(cumsum_rooms), Freq = cumsum_rooms)

# speicher in einem data frame fuer die visualisierung
df_rooms <- data.frame(table_rooms)

# einige ausreisser nach oben
ggplot(miete) + geom_boxplot(aes(y = rooms))
```



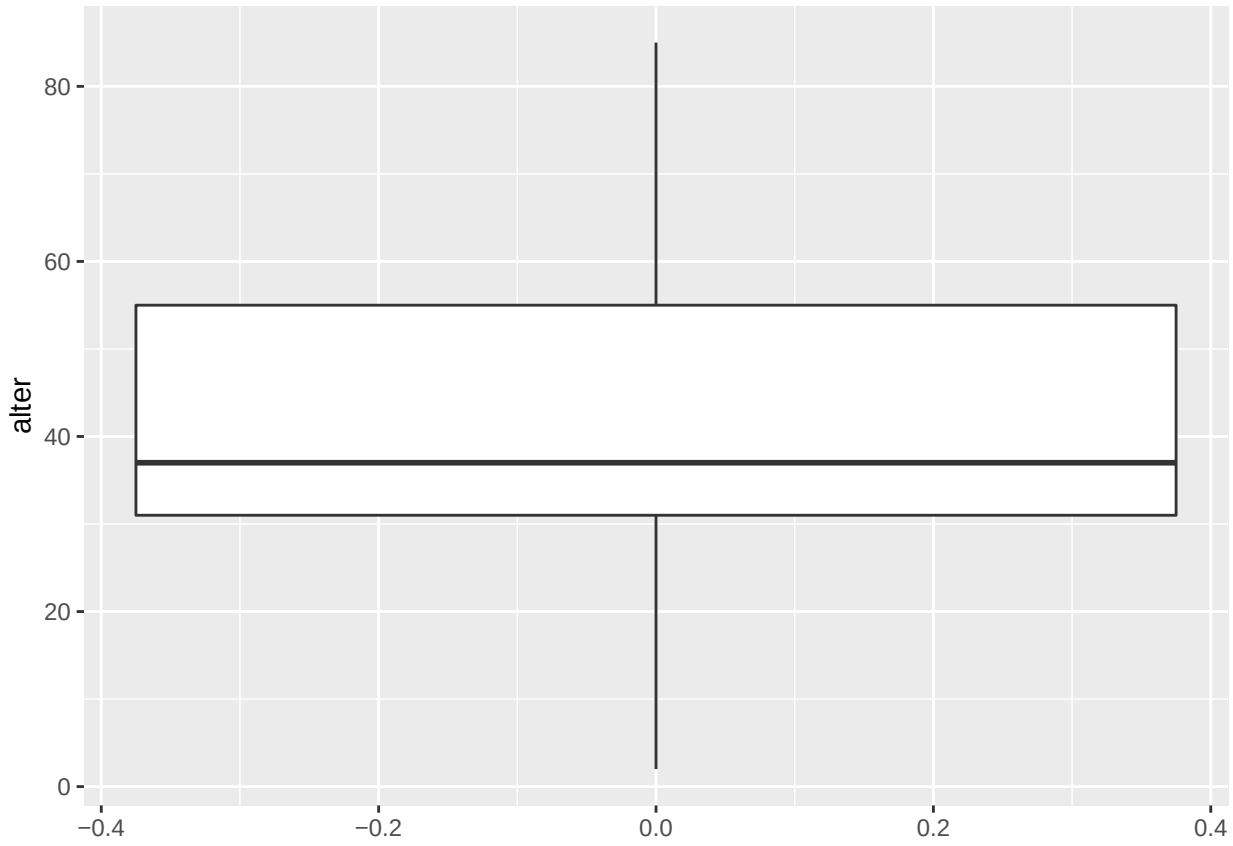

```r
# boxplot(wf)

# histogram
# ggplot(miete) + geom_histogram(aes(x = rooms))

# histogram der kategorien
ggplot(df_rooms, aes(x=rooms, y = Freq)) + geom_bar(stat = "identity") + geom_text(aes(label = Freq), v
       x ="Wohnflaecheintervall", y = "Haeufigkeit")
```



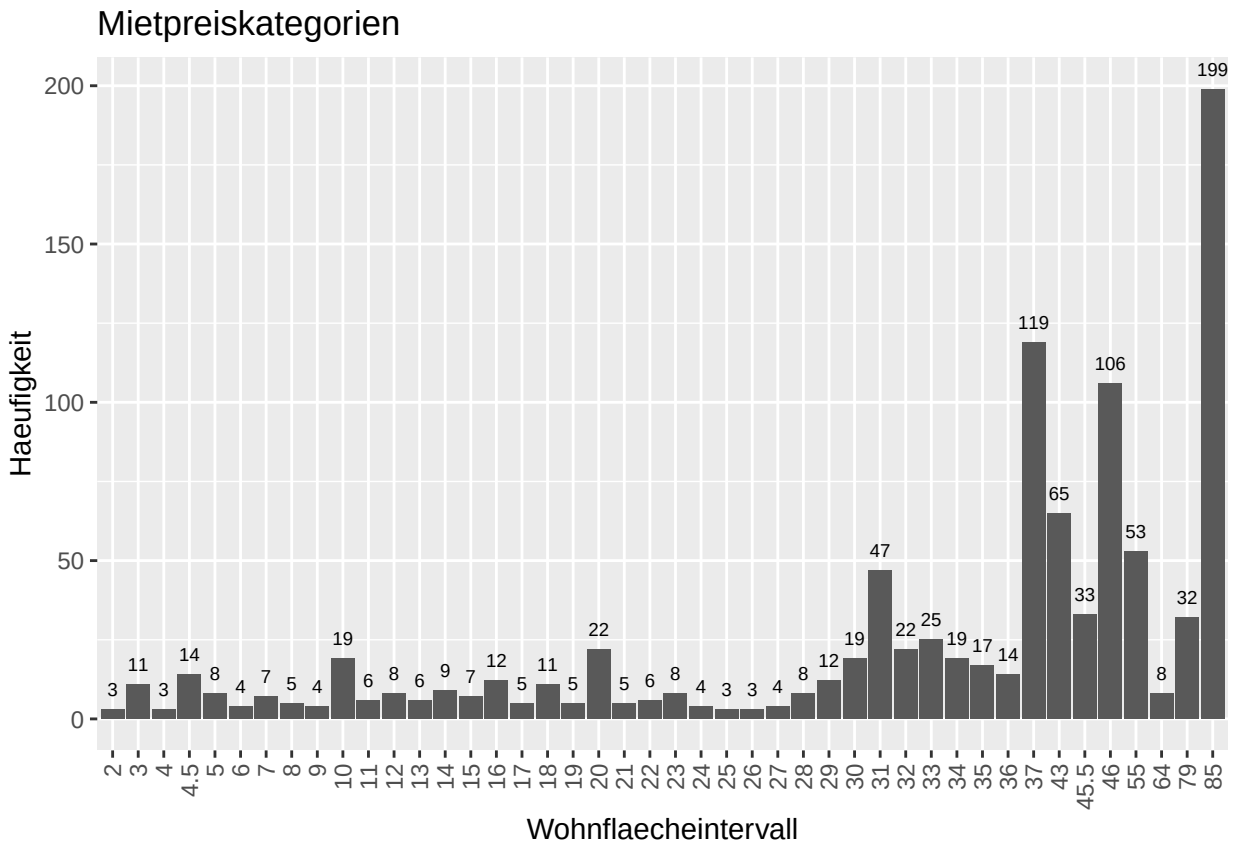

```
# kummulierte haeufigkeiten
df_cumsum_rooms %>% arrange(Freq) %>% mutate(name = factor(Cat, levels=names(cumsum_rooms))) %>% ggplot
```



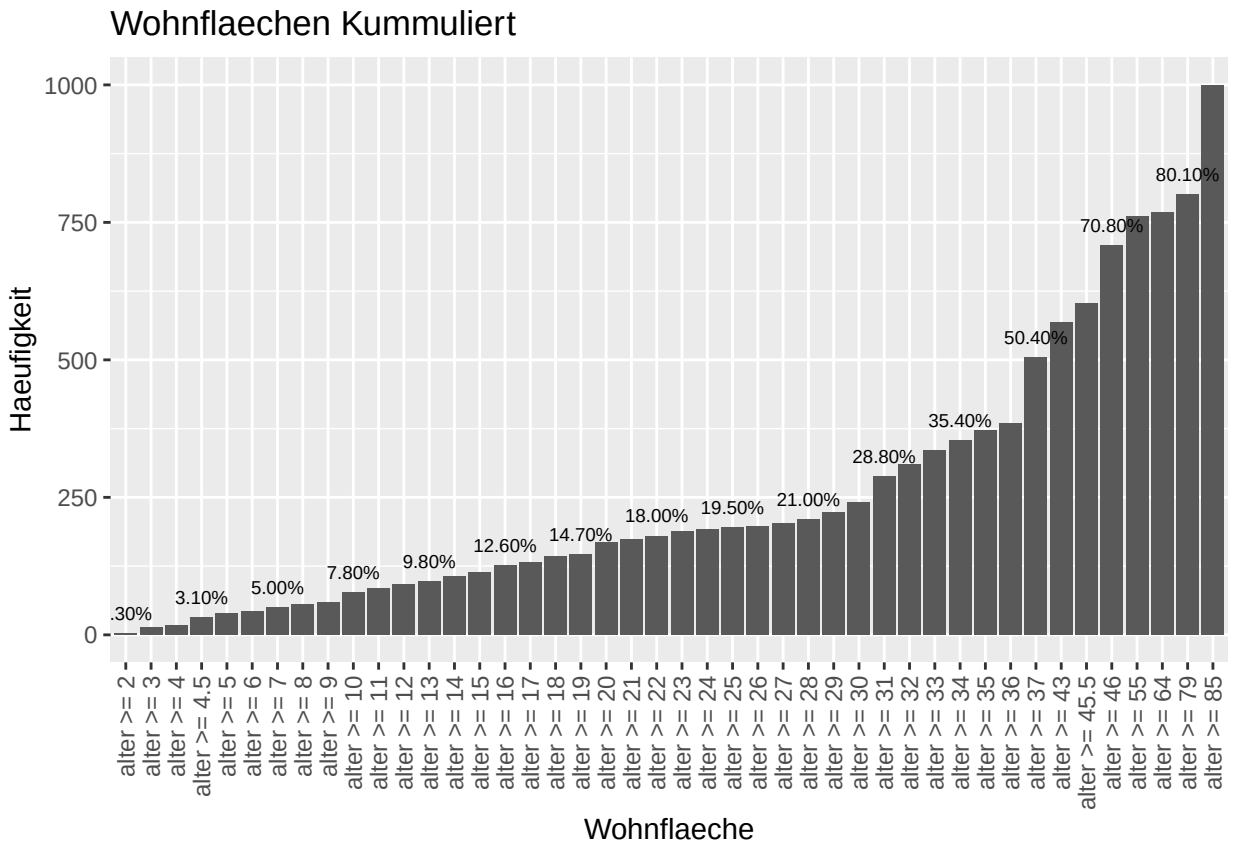

```
# haeufigkeiten der kategorien
table_alter <- table(alter)

# cusum
cumsum_alter <- cumsum(table_alter)
names(cumsum_alter) <- paste("alter >=", names(table_alter))
df_cumsum_alter <- data.frame(Cat = names(cumsum_alter), Freq = cumsum_alter)

# speicher in einem data frame fuer die visualisierung
df_alter <- data.frame(table_alter)

# einige ausreisser nach oben
ggplot(miete) + geom_boxplot(aes(y = alter))
```




```
# boxplot(wf)

# histogram
# ggplot(miete) + geom_histogram(aes(x = alter))

# histogram der kategorien
ggplot(df_alter, aes(x=alter, y = Freq)) + geom_bar(stat = "identity") + geom_text(aes(label = Freq), v
        x ="Wohnflaecheintervall", y = "Haeufigkeit") +
  scale_x_discrete(guide = guide_axis(angle = 90))
```



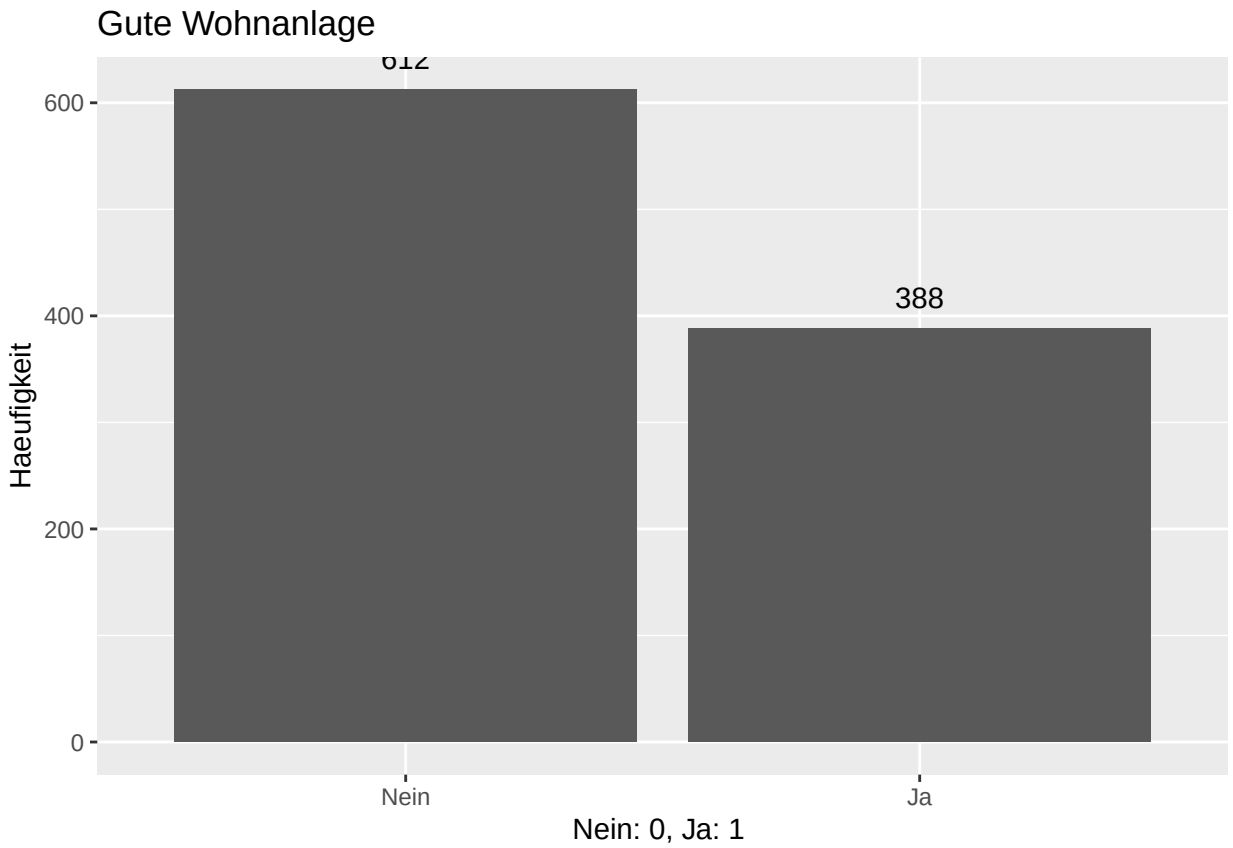

```
# kummulierte haeufigkeiten
df_cumsum_alter %>% arrange(Freq) %>% mutate(name = factor(Cat, levels=names(cumsum_alter))) %>% ggplot
  scale_x_discrete(guide = guide_axis(angle = 90))
```



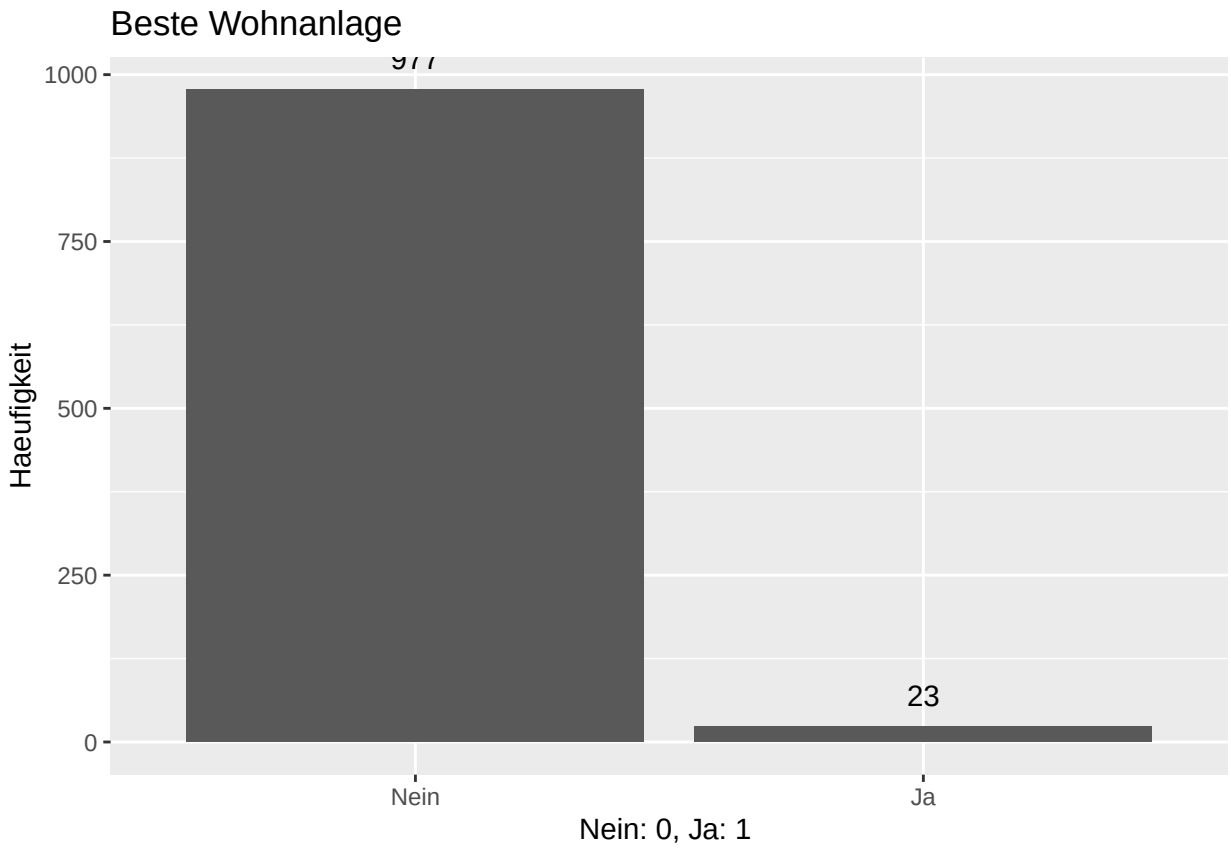

```
names(miete)
```

```
##  [1] "nm"       "wfl"      "rooms"    "bj"       "wohngut"  "wohnbest"
##  [7] "ww0"      "zh0"      "badkach0" "badextra" "kueche"
```

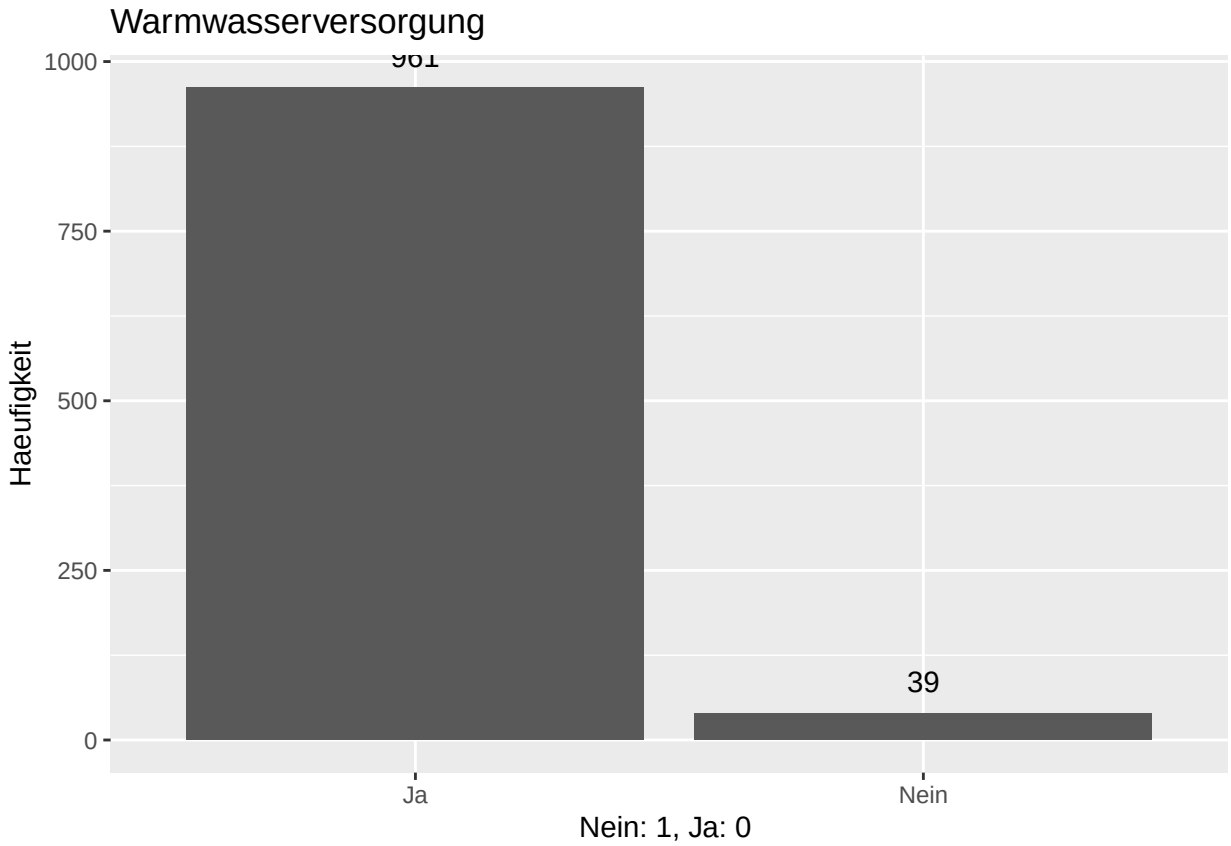
```
table_wg <- table(wohngut)
table_wb <- table(wohnbest)
table_ww0 <- table(ww0)
table_zh0 <- table(zh0)
table_badkach0 <- table(badkach0)
table_badextra <- table(badextra)
table_kueche <- table(kueche)

ggplot(data.frame(table_wg), aes(x = wohngut, y = Freq)) + geom_bar(stat = "identity") + geom_text(label
```



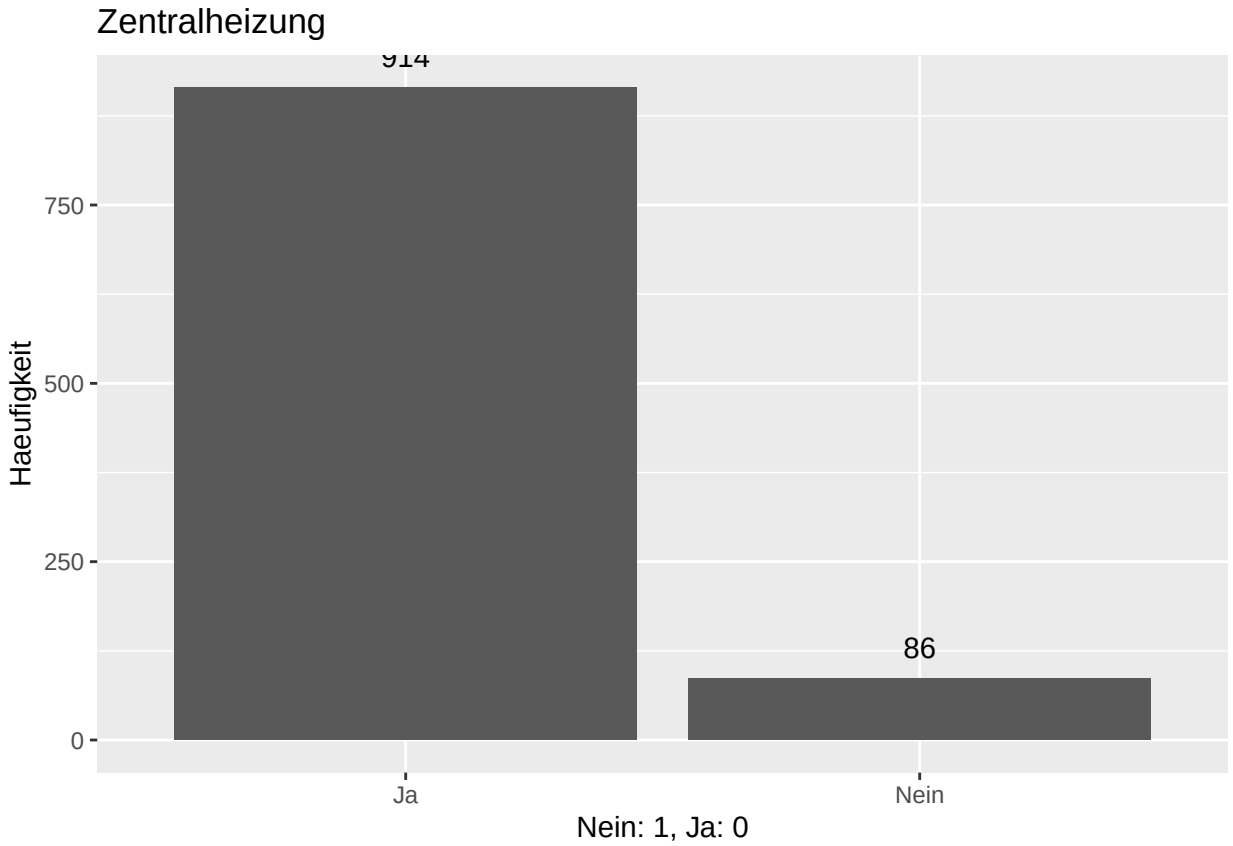

```
ggplot(data.frame(table_wb), aes(x = wohnbest, y = Freq)) + geom_bar(stat = "identity") + geom_text(labe
```



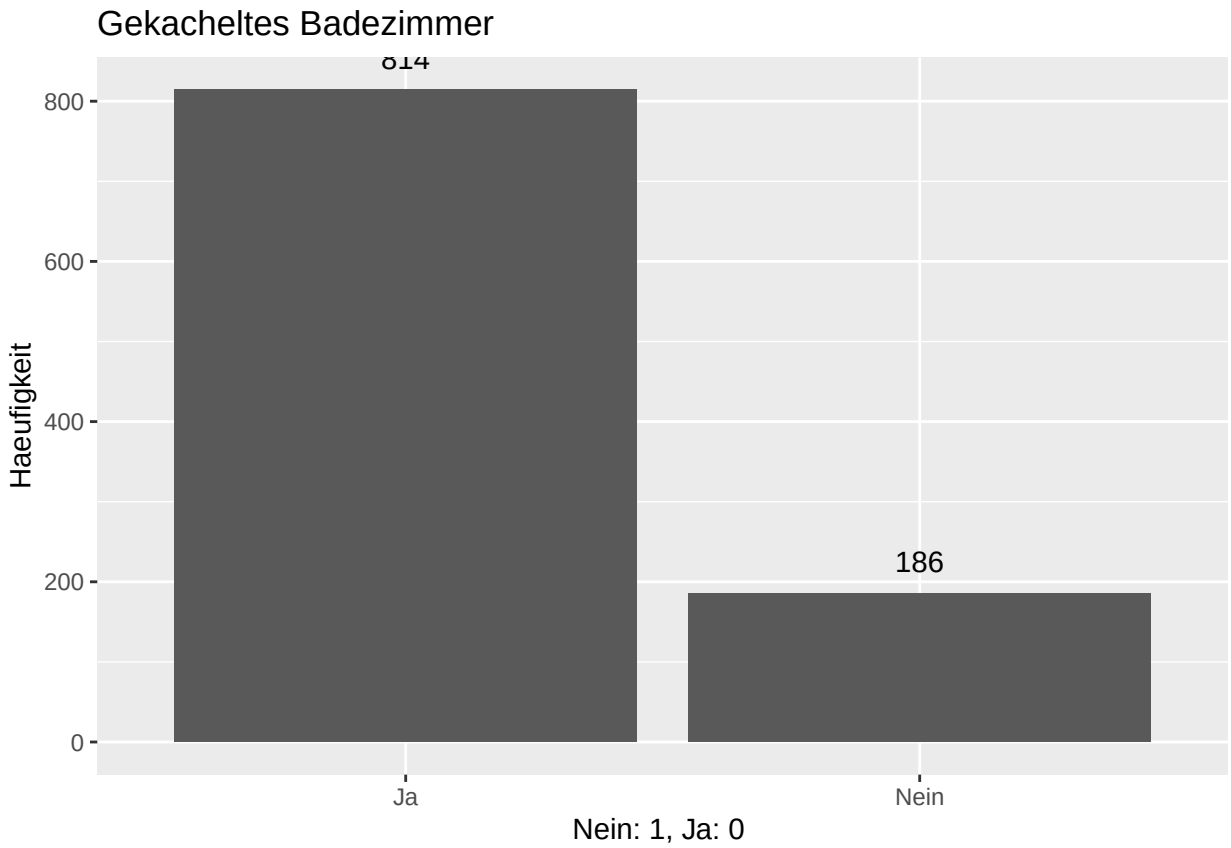

```
ggplot(data.frame(table_ww0), aes(x = ww0, y = Freq)) + geom_bar(stat = "identity") + geom_text(label =
```



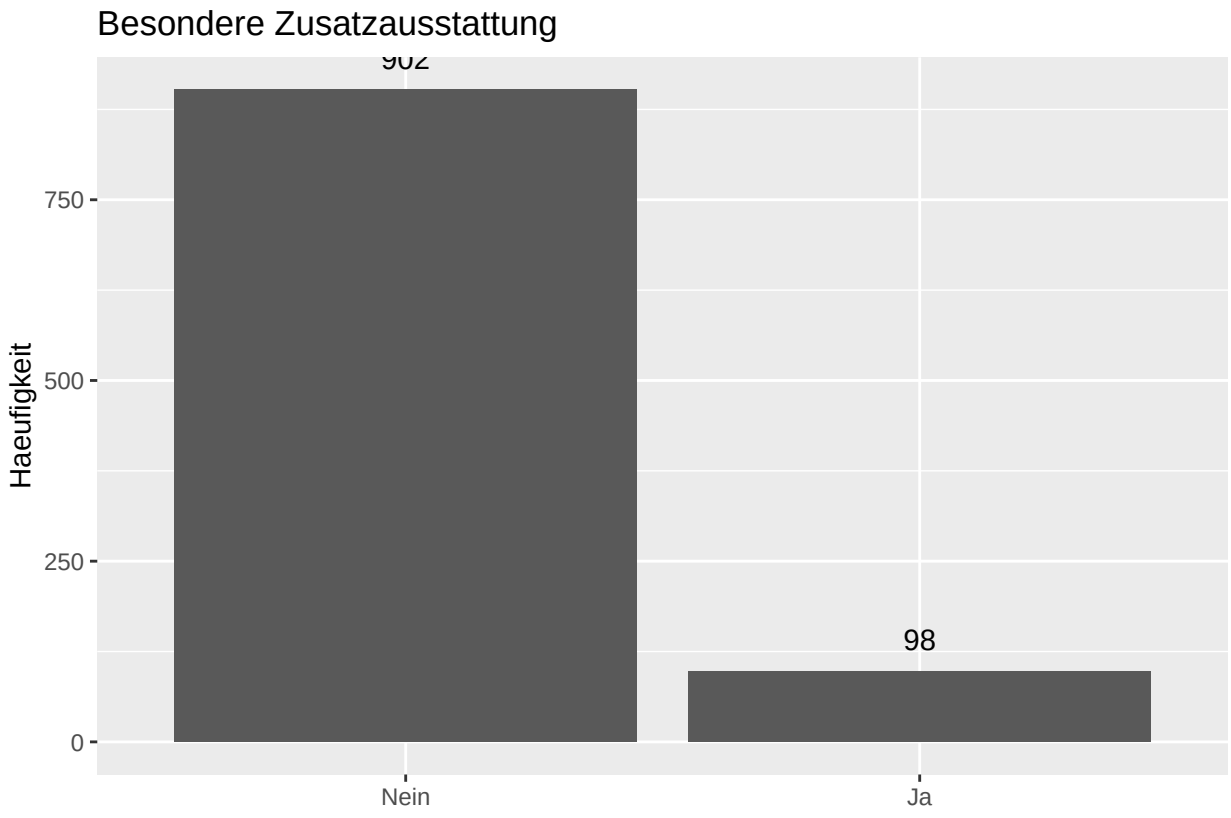

```
ggplot(data.frame(table_zh0), aes(x = zh0, y = Freq)) + geom_bar(stat = "identity") + geom_text(label =
```



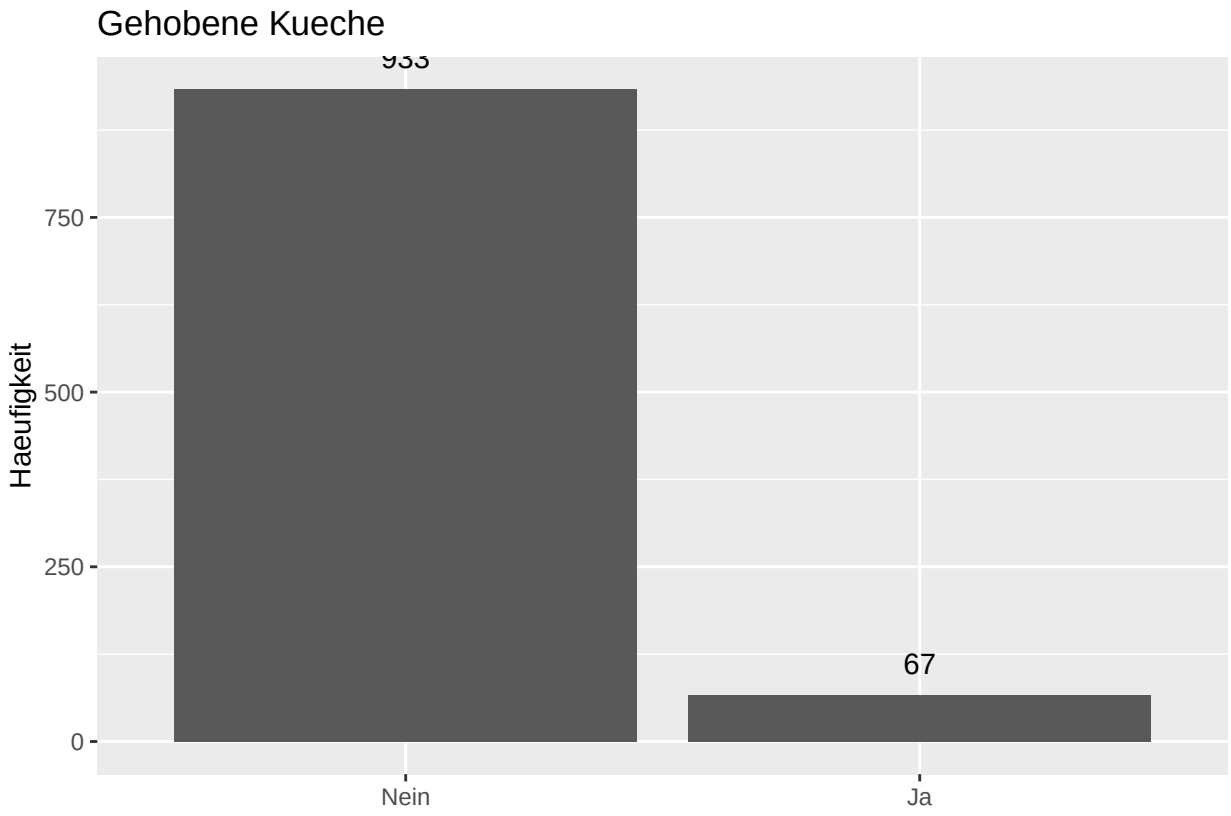

```
ggplot(data.frame(table_badkach0), aes(x = badkach0, y = Freq)) + geom_bar(stat = "identity") + geom_te
```




```
ggplot(data.frame(table_badextra), aes(x = badextra, y = Freq)) + geom_bar(stat = "identity") + geom_te
```



```
ggplot(data.frame(table_kueche), aes(x = kueche, y = Freq)) + geom_bar(stat = "identity") + geom_text(la
```

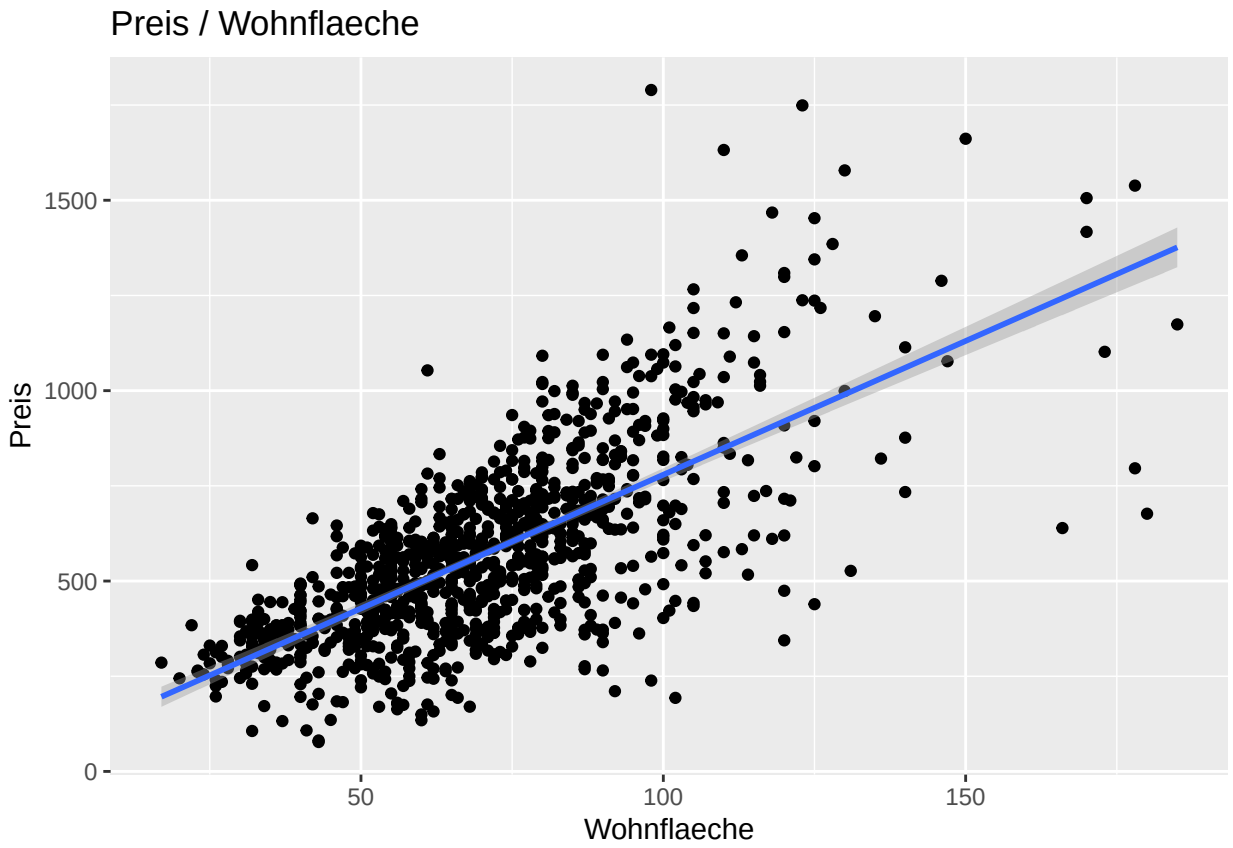



```
names(miete)
```

```
##  [1] "nm"       "wfl"      "rooms"    "bj"       "wohngut"  "wohnbest"
##  [7] "ww0"      "zh0"      "badkach0" "badextra" "kueche"
```

```
# bivariat
# wohnflaeche
ggplot(miete, aes(x = wfl, y = nm)) + geom_point() + geom_smooth(method = "lm") + labs(title="Preis / W
```

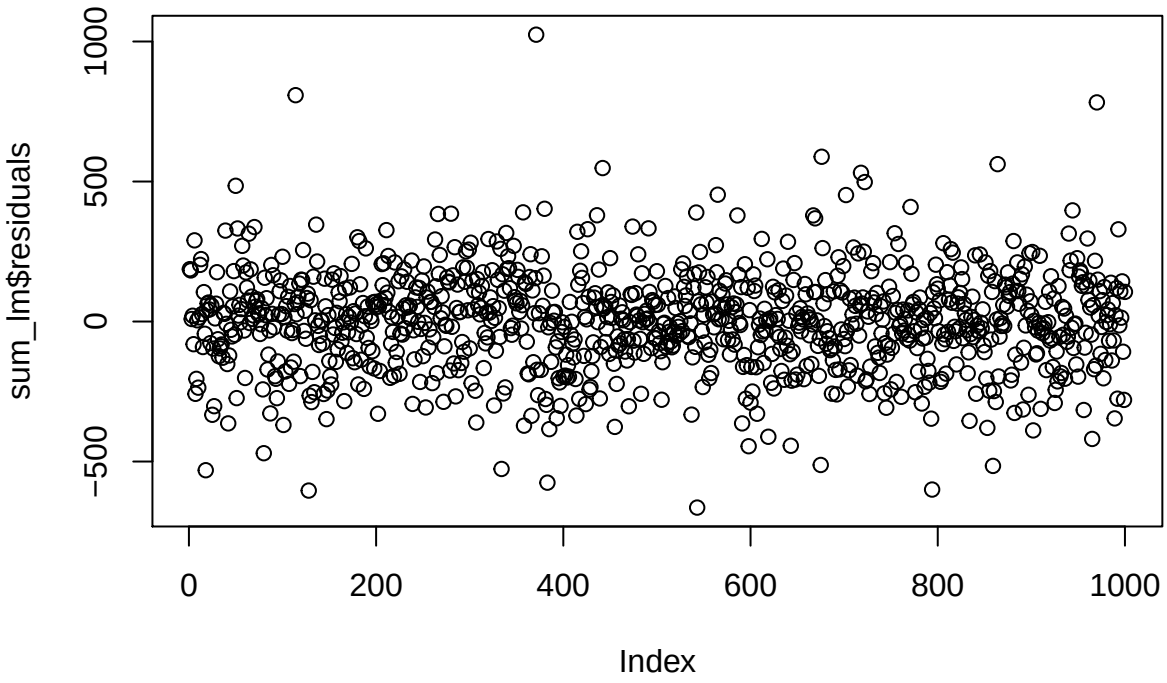
```
## `geom_smooth()` using formula 'y ~ x'
```



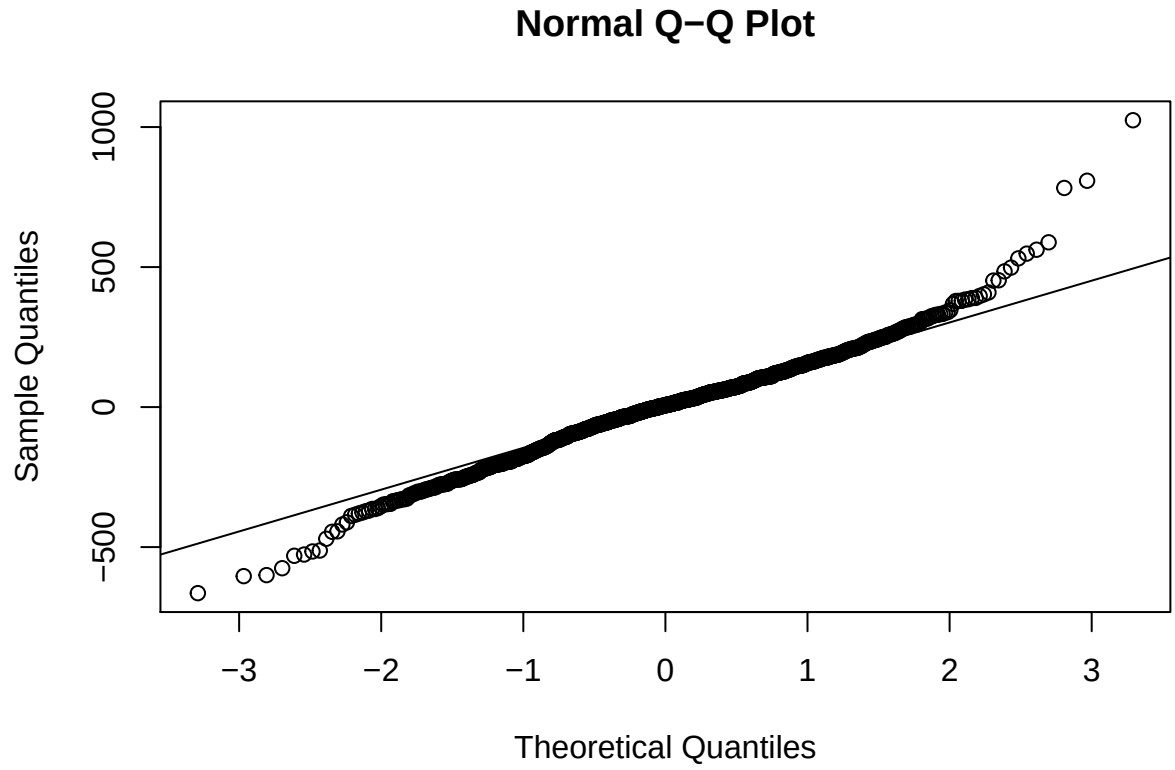

```
sum_lm <- summary(lm(nm ~ wfl))
shapiro.test(sum_lm$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sum_lm$residuals
## W = 0.9793, p-value = 0.00000000009947
```

```
plot(sum_lm$residuals)
```

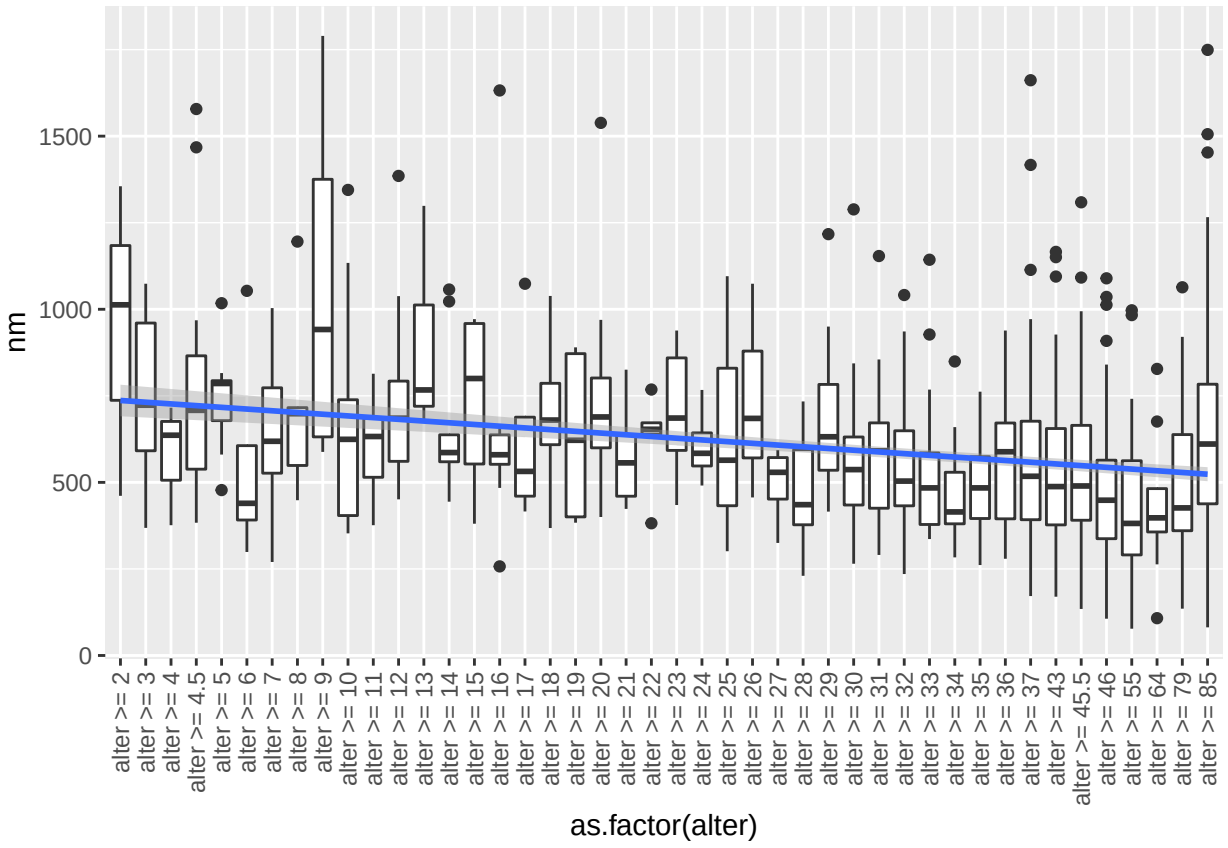



```
qqnorm(sum_lm$residuals)
qqline(sum_lm$residuals)
```




```
# Alter
ggplot(miete, aes(x = as.factor(alter), y = nm)) + geom_boxplot() + geom_smooth(method = "lm", se=TRUE,
  scale_x_discrete(labels = df_cumsum_alter$Cat, guide = guide_axis(angle = 90))
```

```
## `geom_smooth()` using formula 'y ~ x'
```

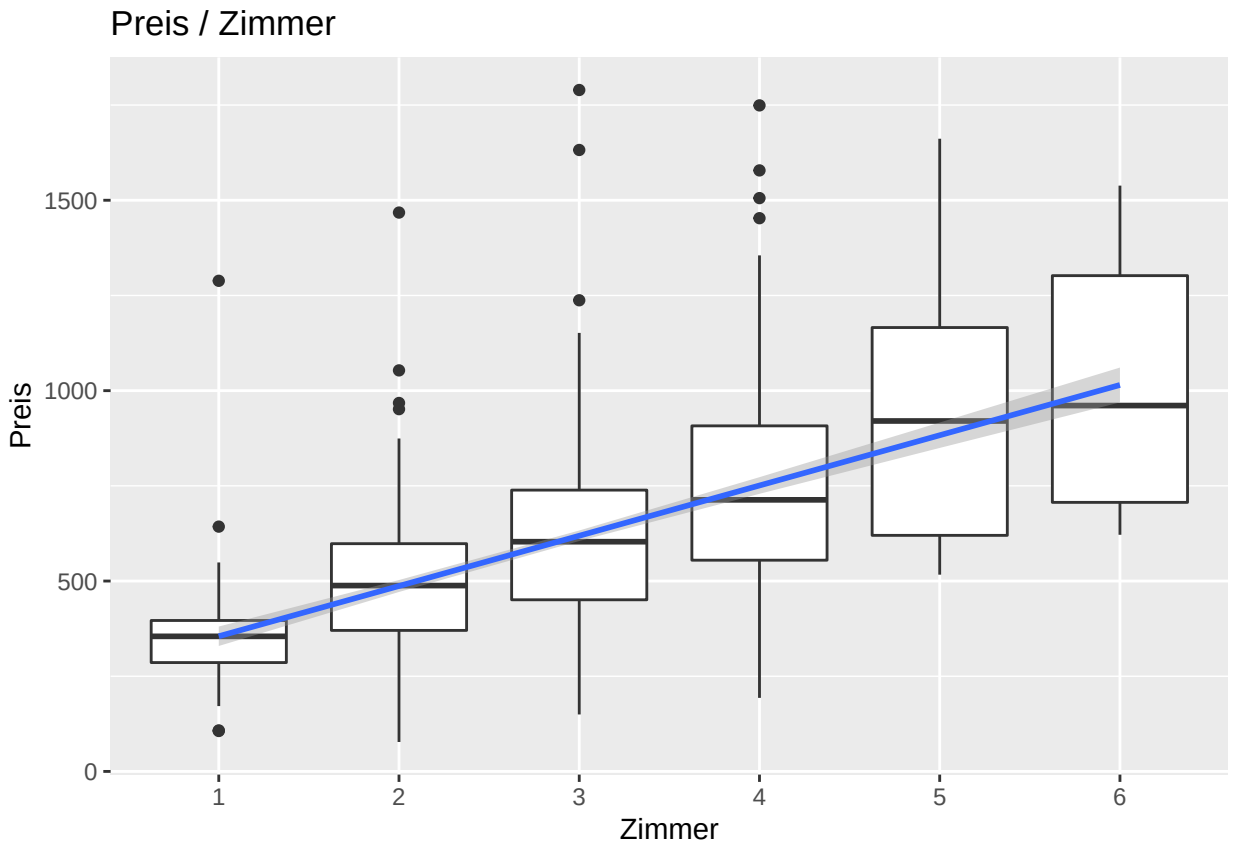



```
summary(lm(nm ~ alter))
```

```
##
## Call:
## lm(formula = nm ~ alter)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -488.77 -183.83  -33.39  114.70 1209.61
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 614.7494    16.2186  37.904  < 2e-16 ***
## alter        -0.8848     0.3155  -2.805  0.00514 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 246.7 on 998 degrees of freedom
## Multiple R-squared:  0.00782,    Adjusted R-squared:  0.006825
## F-statistic: 7.865 on 1 and 998 DF,  p-value: 0.005137
```

```
# Zimmer
ggplot(miete, aes(x = as.factor(rooms), y = nm)) + geom_boxplot() + geom_smooth(method = "lm", se=TRUE,
```

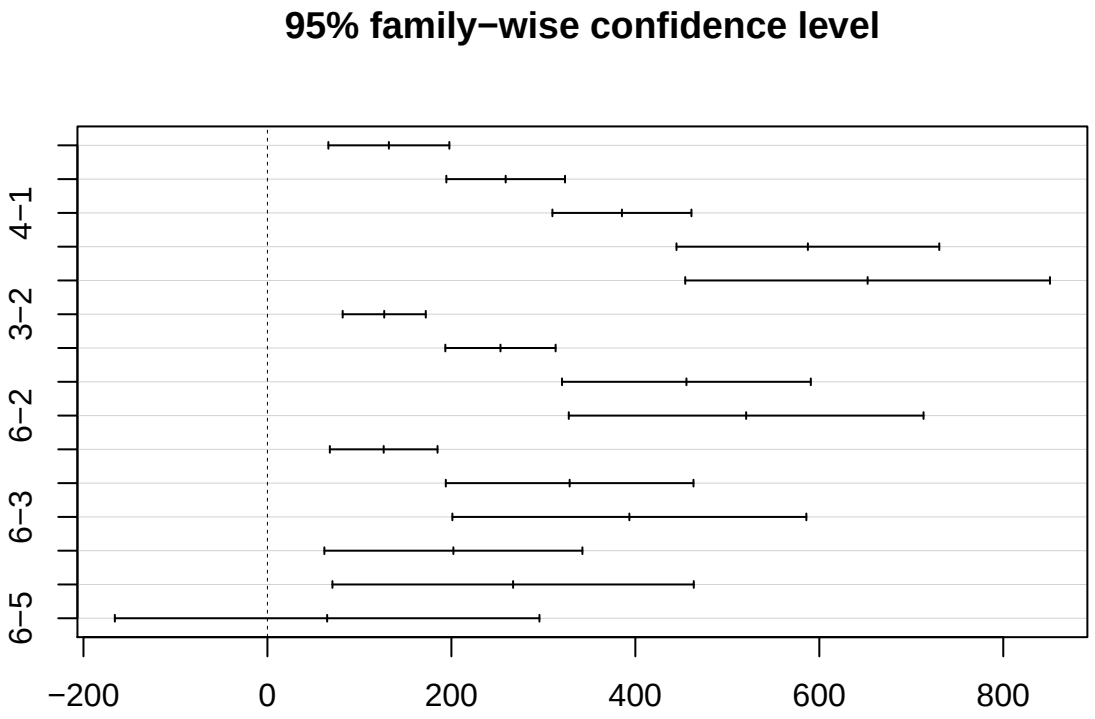
```
## `geom_smooth()` using formula 'y ~ x'
```




```
summary(aov(nm ~ factor(rooms)))
```

```
##                Df   Sum Sq Mean Sq F value Pr(>F)
## factor(rooms)   5 17206255 3441251   77.74 <2e-16 ***
## Residuals     994 43999054   44265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
avg_rent_room <- aggregate(nm, list(rooms), mean)
plot(TukeyHSD(aov(nm ~ factor(rooms))))
```




```
# binaeren
ggplot(miete, aes(x = as.factor(wohngut), y = nm)) + geom_boxplot() + labs(title="Preis / Gute Wohnanlag
```

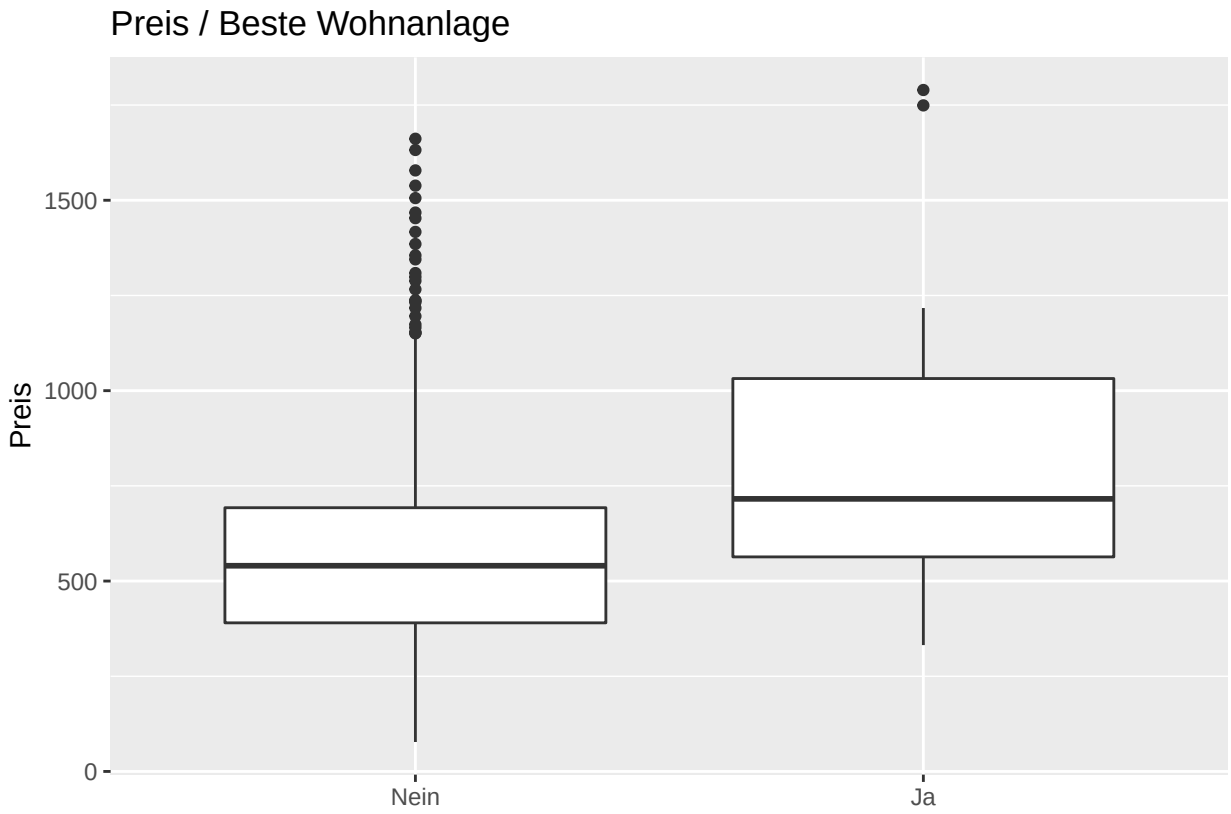
```r
summary(aov(nm ~ wohngut))
```

```
##              Df   Sum Sq Mean Sq F value      Pr(>F)
## wohngut       1  1471706 1471706   24.59 0.000000834 ***
## Residuals   998 59733603   59853
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(lm(nm ~ wohngut))
```

```
##
## Call:
## lm(formula = nm ~ wohngut)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -467.01 -171.09  -27.06  116.11 1245.23
##
## Coefficients:
##             Estimate Std. Error t value    Pr(>|t|)
## (Intercept)  544.325      9.889  55.041     < 2e-16 ***
## wohngut       78.726     15.876   4.959 0.000000834 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 244.6 on 998 degrees of freedom
## Multiple R-squared:  0.02405,    Adjusted R-squared:  0.02307
## F-statistic: 24.59 on 1 and 998 DF,  p-value: 0.0000008337
```

```r
# beste wohnanlage
ggplot(miete, aes(x = as.factor(wohnbest), y = nm)) + geom_boxplot() + labs(title="Preis / Beste Wohnan
```



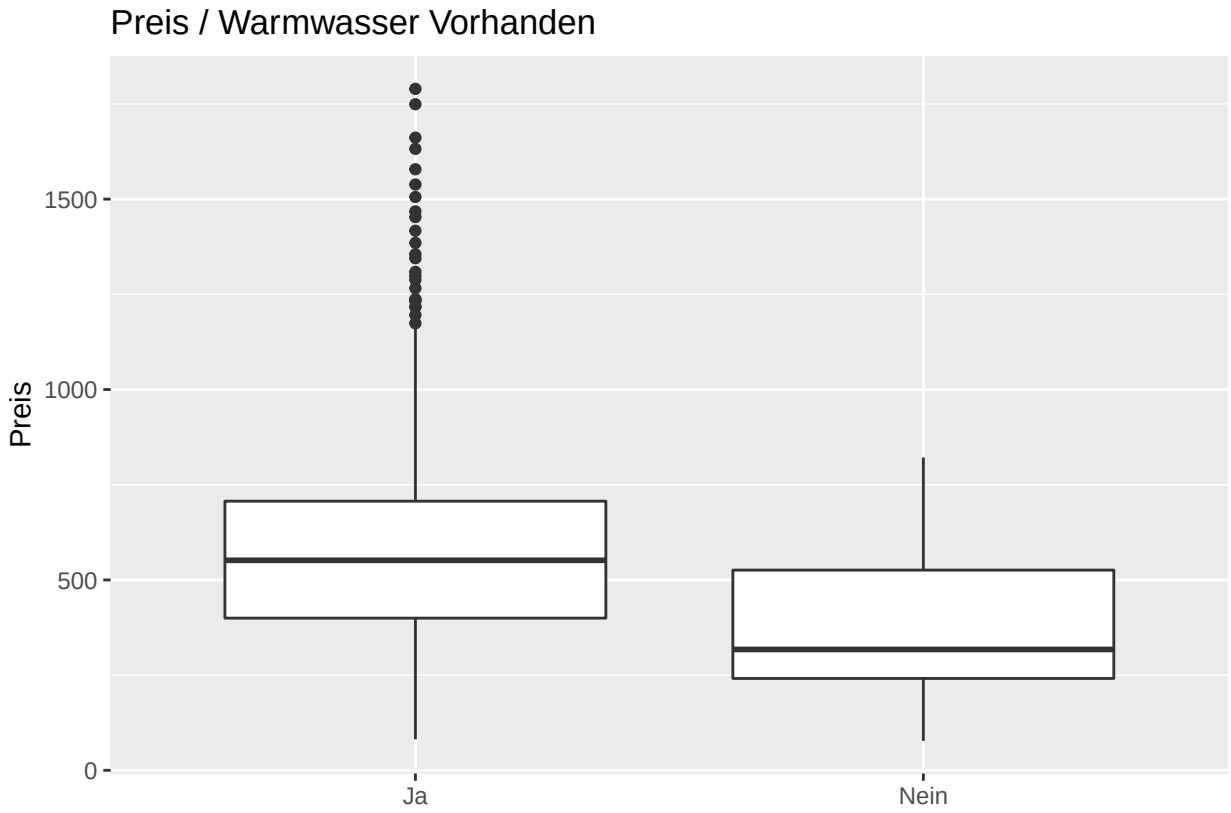

```
summary(aov(nm ~ wohnbest))
```

```
##              Df   Sum Sq Mean Sq F value      Pr(>F)
## wohnbest      1  1525903 1525903   25.52 0.000000521 ***
## Residuals   998 59679406   59799
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm(nm ~ wohnbest))
```

```
##
## Call:
## lm(formula = nm ~ wohnbest)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -497.59 -180.31  -29.56  126.21 1092.67
##
## Coefficients:
##             Estimate Std. Error t value    Pr(>|t|)
## (Intercept)  568.877      7.823  72.714     < 2e-16 ***
## wohnbest     260.587     51.586   5.051 0.000000521 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 244.5 on 998 degrees of freedom
## Multiple R-squared:  0.02493,    Adjusted R-squared:  0.02395
## F-statistic: 25.52 on 1 and 998 DF,  p-value: 0.0000005212
```

```r
# warmwasser
ggplot(miete, aes(x = as.factor(ww0), y = nm)) + geom_boxplot() + labs(title="Preis / Warmwasser Vorhan
```




```r
summary(aov(nm ~ ww0))
```

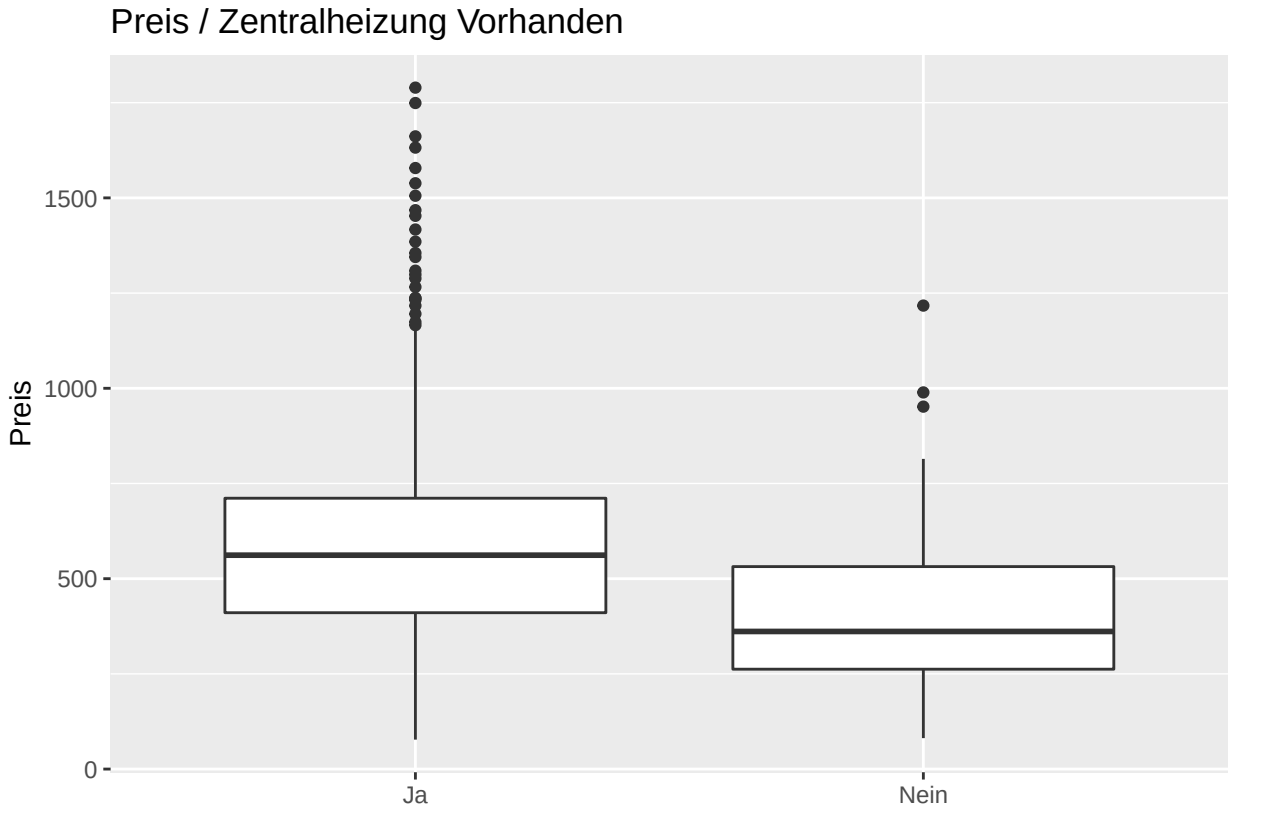
```
##              Df   Sum Sq Mean Sq F value      Pr(>F)
## ww0           1  1527877 1527877   25.55 0.000000512 ***
## Residuals   998 59677432   59797
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(lm(nm ~ ww0))
```

```
##
## Call:
## lm(formula = nm ~ ww0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -501.46 -180.34  -32.49  124.94 1206.81
##
## Coefficients:
##             Estimate Std. Error t value    Pr(>|t|)
## (Intercept)  582.745      7.888  73.875     < 2e-16 ***
## ww0         -201.906     39.943  -5.055 0.000000512 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 244.5 on 998 degrees of freedom
```

```
## Multiple R-squared:  0.02496,    Adjusted R-squared:  0.02399
## F-statistic: 25.55 on 1 and 998 DF,  p-value: 0.0000005123
```

```
# zentralheizung
ggplot(miete, aes(x = as.factor(zh0), y = nm)) + geom_boxplot() + labs(title="Preis / Zentralheizung Vo
```




```
summary(aov(nm ~ zh0))
```

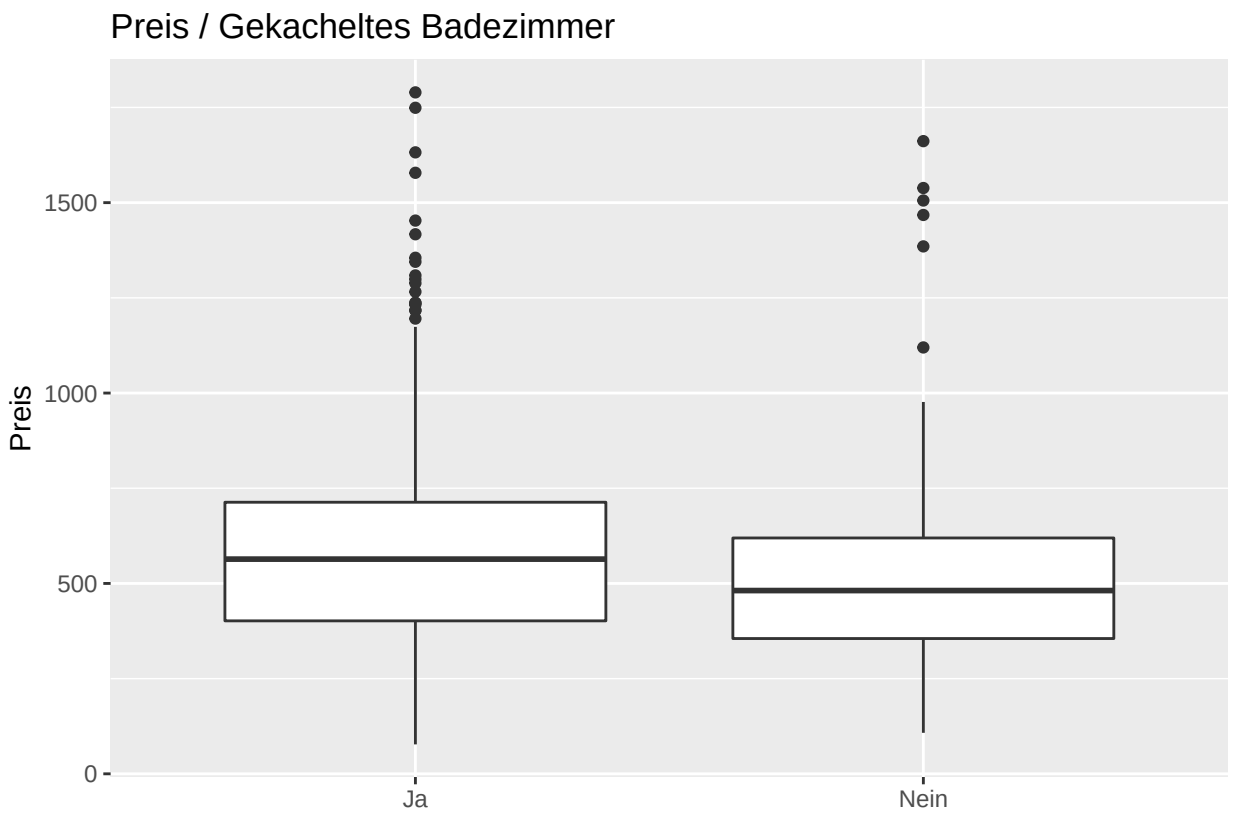
```
##              Df   Sum Sq Mean Sq F value         Pr(>F)
## zh0           1  2497857 2497857   42.46 0.000000000114 ***
## Residuals   998 58707451   58825
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm(nm ~ zh0))
```

```
##
## Call:
## lm(formula = nm ~ zh0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -512.89 -174.61  -29.98  120.87 1199.35
##
## Coefficients:
##             Estimate Std. Error t value       Pr(>|t|)
## (Intercept)  590.201      8.022  73.568        < 2e-16 ***
## zh0         -178.263     27.356  -6.516 0.000000000114 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 242.5 on 998 degrees of freedom
## Multiple R-squared:  0.04081,    Adjusted R-squared:  0.03985
## F-statistic: 42.46 on 1 and 998 DF,  p-value: 0.0000000001141
```

```
# gekacheltes badezimmer
ggplot(miete, aes(x = as.factor(badkach0), y = nm)) + geom_boxplot() + labs(title="Preis / Gekacheltes
```




```
summary(aov(nm ~ badkach0))
```

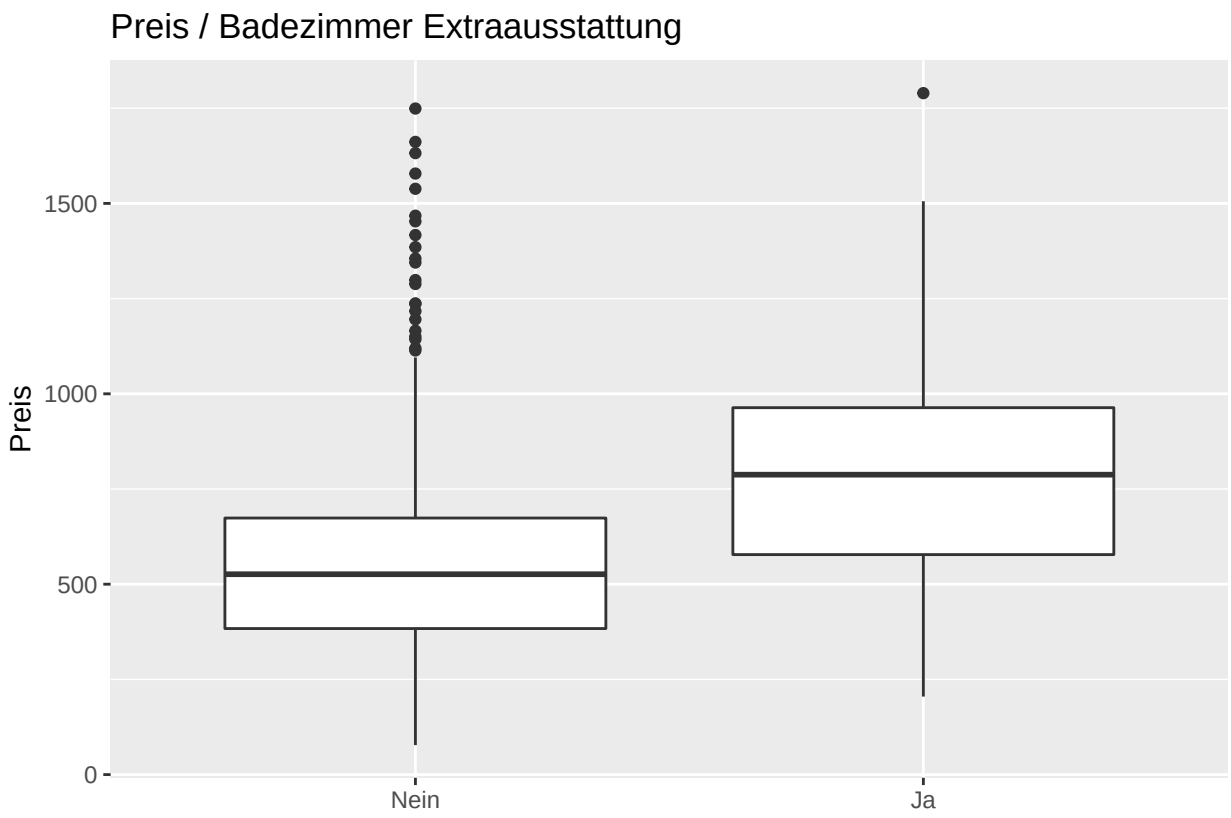
```
##              Df   Sum Sq Mean Sq F value   Pr(>F)
## badkach0      1   849011  849011   14.04 0.000189 ***
## Residuals   998 60356297   60477
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm(nm ~ badkach0))
```

```
##
## Call:
## lm(formula = nm ~ badkach0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -511.49 -180.16  -26.37  122.30 1200.75
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   588.80       8.62  68.310  < 2e-16 ***
## badkach0      -74.88      19.99  -3.747 0.000189 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 245.9 on 998 degrees of freedom
## Multiple R-squared:  0.01387,    Adjusted R-squared:  0.01288
## F-statistic: 14.04 on 1 and 998 DF,  p-value: 0.0001894
```

```
# badezimmer extraausstattung
ggplot(miete, aes(x = as.factor(badextra), y = nm)) + geom_boxplot() + labs(title="Preis / Badezimmer E
```




```
summary(aov(nm ~ badextra))
```

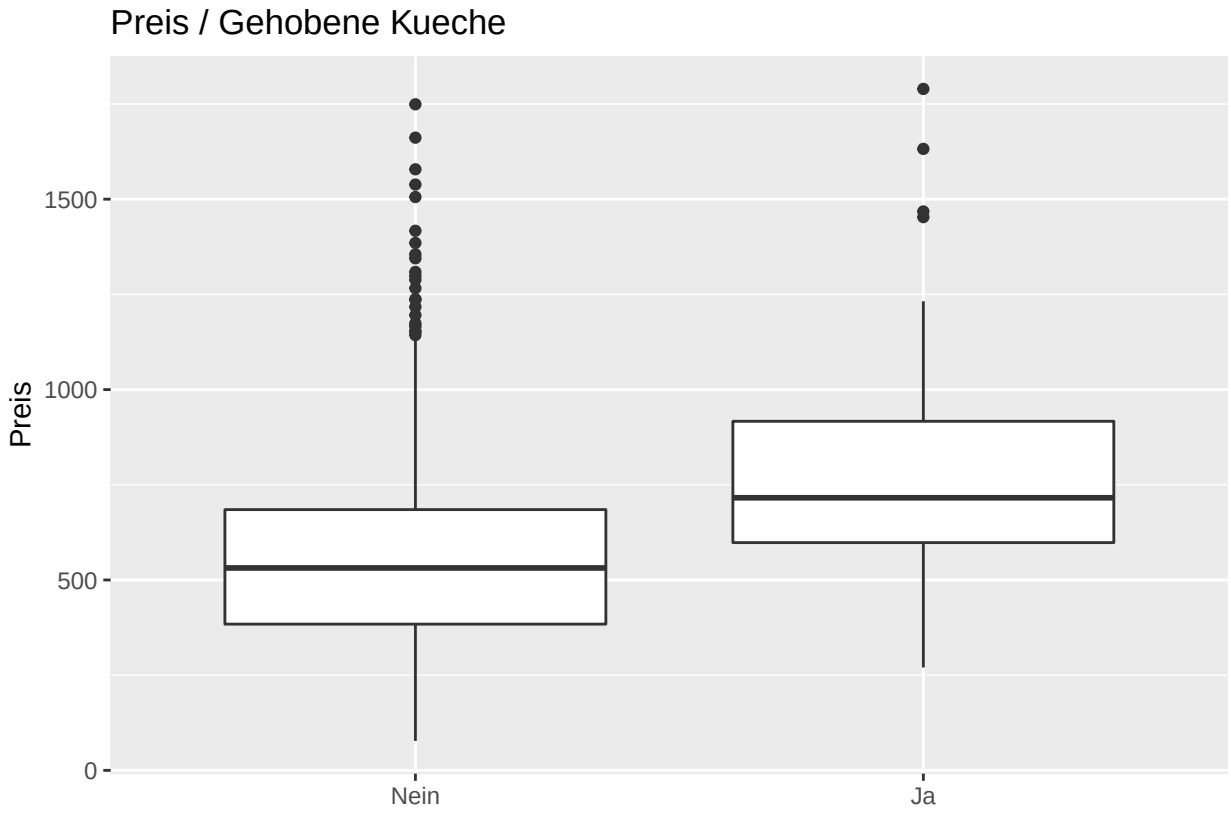
```
##              Df   Sum Sq Mean Sq F value Pr(>F)
## badextra      1  4464893 4464893   78.53 <2e-16 ***
## Residuals   998 56740415   56854
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm(nm ~ badextra))
```

```
##
## Call:
## lm(formula = nm ~ badextra)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -572.74 -169.40  -24.95  123.86 1196.30
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  552.845      7.939  69.635   <2e-16 ***
## badextra     224.745     25.361   8.862   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 238.4 on 998 degrees of freedom
## Multiple R-squared:  0.07295,    Adjusted R-squared:  0.07202
## F-statistic: 78.53 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
# gehobene kueche
ggplot(miete, aes(x = as.factor(kueche), y = nm)) + geom_boxplot() + labs(title="Preis / Gehobene Kueche
```




```
summary(aov(nm ~ kueche))
```

```
##              Df   Sum Sq Mean Sq F value           Pr(>F)
## kueche        1  2896574 2896574   49.58 0.00000000000355 ***
## Residuals   998 58308734   58426
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm(nm ~ kueche))
```

```
##
## Call:
## lm(formula = nm ~ kueche)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -505.64 -176.43  -33.01  124.36 1188.70
##
```

```
## Coefficients:
##             Estimate Std. Error t value         Pr(>|t|)
## (Intercept)  560.448      7.913  70.823          < 2e-16 ***
## kueche       215.260     30.572   7.041 0.00000000000355 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 241.7 on 998 degrees of freedom
## Multiple R-squared:  0.04733,    Adjusted R-squared:  0.04637
## F-statistic: 49.58 on 1 and 998 DF,  p-value: 0.000000000003546
```

```
names(miete)
```

```
##  [1] "nm"       "wfl"      "rooms"    "bj"       "wohngut"  "wohnbest"
##  [7] "ww0"      "zh0"      "badkach0" "badextra" "kueche"
```

```
# korrelation wfl / rooms
cor(wfl, rooms)
```

```
## [1] 0.8394882
```

```
phi <- function(tab){
        (tab[1,1] * tab[2,2] - tab[1,2] * tab[2,1]) / sqrt(rowSums(tab)[1] * rowSums(tab)[2] *  colSums
}

n_biv <- length(names(miete)[-(1:4)])
bivarite_phi_coef <- matrix(0, ncol = n_biv, nrow = n_biv)

for(i in 1:n_biv){
        for(i in 1:n_biv){

        }
}

tab <- table(wohngut, wohnbest)
phi(tab)
```

```
##          0
## -0.1221678
```