## Crashkurs R

Wir arbeiten mit dem Paket ‘mosaic’, darum laden wir es zu Beginn. Nützlich ist auch das Paket ‘here’, welches wir daher umgehend auch laden:

library(mosaic) #   
library(here) #

### Die zentrale Idee von ‘mosaic’

analysiere( y # ggfs. abhängige Variable  
 ~ x # unabhängige Variable(n)  
 | z, # ggfs. bedingende (gruppierende) Variable(n)  
 Optionen, # ggfs. weitere Optionen  
 data = daten ) # Datensatz

analysiere(): Was soll R tun?[[1]](#footnote-20)

*Hinweis*: unter macOS: ~: alt+n oder option+n, |: alt+7 oder option+7

### Zentrale Fragen die SIE sich stellen sollten!

1. Was soll der Computer für mich tun? –> meineanalyse
2. Was muss der Computer dafür wissen? –> Parameter (....)

meineanalyse(meiny ~ meinx, data = meinedaten)

### Vorbereitung Analyse

[Dick De Veaux: How much is a Fireplace Worth?](http://community.amstat.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=288c3e05-1ba5-450d-8ec8-62629b876557&forceDialog=0)[[2]](#footnote-25)

* Preis: Preis in .
* Wohnflaeche: Wohnfläche in .
* Alter: Alter der Immobilie in Jahren.
* Klimaanlage: Inwieweit eine (zentrale) Klimaanlage vorhanden ist.
* Kamin: Inwieweit ein Kamin vorhanden ist.
* Heizung: Heizsystem: Gas, Strom oder Öl.

# Paket laden  
library(mosaic)  
  
# URL für die Datentabelle  
daten\_url <- "http://statistix.org/Data/SaratogaHouses.csv"  
  
# Daten in R aus derCSV-Datei einlesen  
Houses <- read.csv2(daten\_url)

### Inspizieren der Datentabelle

inspect(Houses)

##   
## categorical variables:   
## name class levels n missing  
## 1 Klimaanlage character 2 1728 0  
## 2 Kamin character 2 1728 0  
## 3 Heizung character 3 1728 0  
## distribution  
## 1 Nein (63.3%), Ja (36.7%)   
## 2 Ja (57.2%), Nein (42.8%)   
## 3 Gas (69.3%), Strom (18.2%) ...   
##   
## quantitative variables:   
## name class min Q1 median Q3  
## ...1 Preis integer 5000.0000 145000.0000 189900.0000 259000.0000  
## ...2 Wohnflaeche numeric 57.2278 120.7729 151.8488 198.6018  
## ...3 Alter integer 0.0000 13.0000 19.0000 34.0000  
## max mean sd n missing  
## ...1 775000.0000 211966.70544 98441.39102 1728 0  
## ...2 485.6931 163.04122 57.59342 1728 0  
## ...3 225.0000 27.91609 29.20999 1728 0

### Eine Variable

**meinx: kategorial**

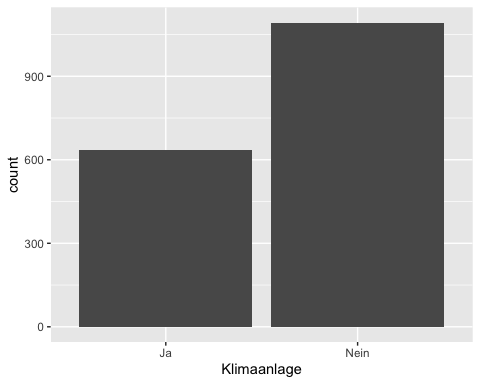
# Säulendiagramm  
gf\_bar( ~ Klimaanlage,   
 data = Houses)  
# Tabelle  
tally( ~ Klimaanlage,   
 data = Houses)  
# Anteil  
prop( ~ Klimaanlage,   
 data = Houses,   
 success = "Ja")

**meinx: numerisch**

# Histogramm  
gf\_histogram( ~ Preis,   
 data = Houses)  
# Kennzahlen  
favstats( ~ Preis,   
 data = Houses)  
# Mittelwert  
mean( ~ Preis,   
 data = Houses)

### Eine kategoriale Variable

# Säulendiagramm  
gf\_bar( ~ Klimaanlage, data = Houses)



# Tabelle  
tally( ~ Klimaanlage, data = Houses)

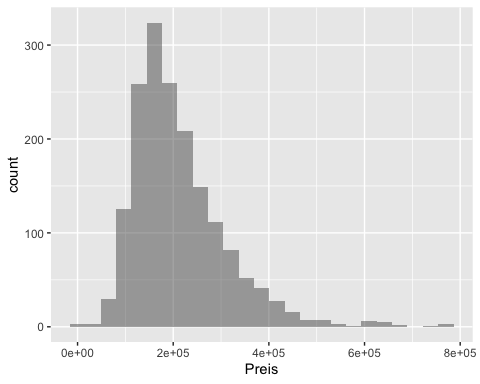
## Klimaanlage  
## Ja Nein   
## 635 1093

# Anteil  
prop( ~ Klimaanlage, success = "Ja", data = Houses)

## prop\_Ja   
## 0.3674769

### Eine numerische Variable

# Histogramm  
gf\_histogram( ~ Preis, data = Houses)



# Kennzahlen  
favstats( ~ Preis, data = Houses)

## min Q1 median Q3 max mean sd n missing  
## 5000 145000 189900 259000 775000 211966.7 98441.39 1728 0

# Mittelwert  
mean( ~ Preis, data = Houses)

## [1] 211966.7

### Zwei Variablen

**meinx, meiny: kategorial**

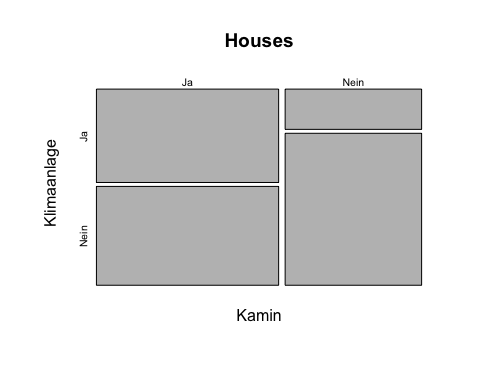
# Mosaikplot  
mosaicplot(Kamin ~ Klimaanlage,   
 data = Houses)  
# Kreuztabelle  
tally(Kamin ~ Klimaanlage,   
 data = Houses)  
# Chi-Quadrat Test  
xchisq.test(Kamin ~ Klimaanlage,   
 data = Houses)

**meinx, meiny: metrisch**

# Streudiagramm  
gf\_point(Preis ~ Wohnflaeche,   
 data = Houses)  
# Korrelation  
cor(Preis ~ Wohnflaeche,   
 data = Houses)  
# Korrelationstest  
cor.test(Preis ~ Wohnflaeche,   
 data = Houses)

### Zwei kategoriale Variablen (I/II)

# Mosaikplot  
mosaicplot(Kamin ~ Klimaanlage, data = Houses)



# Kreuztabelle  
tally(Kamin ~ Klimaanlage, data = Houses)

## Klimaanlage  
## Kamin Ja Nein  
## Ja 480 508  
## Nein 155 585

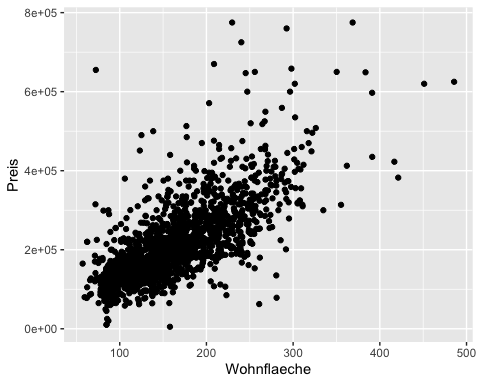
### Zwei kategoriale Variablen (II/II)

# Chi-Quadrat Test  
xchisq.test(Kamin ~ Klimaanlage, data = Houses)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: x  
## X-squared = 137.85, df = 1, p-value < 2.2e-16  
##   
## 480 508   
## (363.07) (624.93)  
## [37.34] [21.69]   
## < 6.14> <-4.68>   
##   
## 155 585   
## (271.93) (468.07)  
## [49.85] [28.96]   
## <-7.09> < 5.40>   
##   
## key:  
## observed  
## (expected)  
## [contribution to X-squared]  
## <Pearson residual>

### Zwei numerische Variablen (I/II)

# Streudiagramm  
gf\_point(Preis ~ Wohnflaeche, data = Houses)



# Korrelation  
cor(Preis ~ Wohnflaeche, data = Houses)

## [1] 0.7123902

### Zwei numerische Variablen (II/II)

# Korrelationstest  
cor.test(Preis ~ Wohnflaeche, data = Houses)

##   
## Pearson's product-moment correlation  
##   
## data: Preis and Wohnflaeche  
## t = 42.173, df = 1726, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.6883589 0.7348595  
## sample estimates:  
## cor   
## 0.7123902

### Zwei Gruppen

**meinx: binär, meiny: kategorial**

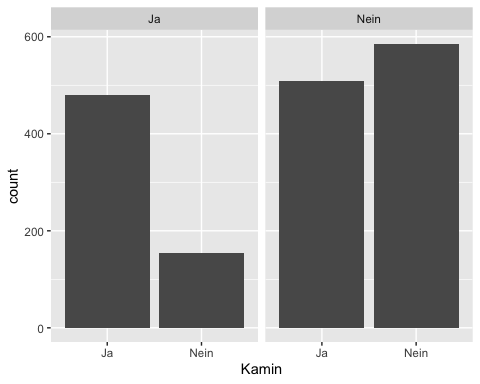
# Säulendiagramm  
gf\_bar( ~ Kamin | Klimaanlage,   
 data = Houses)  
# Anteile  
prop(Kamin ~ Klimaanlage,   
 data = Houses,  
 success = "Ja")  
# Anteilstest  
prop.test(Kamin ~ Klimaanlage,   
 data = Houses,  
 success = "Ja")

**meinx: binär, meiny: numerisch**

# Histogramm  
gf\_histogram( ~ Preis | Kamin,   
 data = Houses)  
# Mittelwerte  
mean(Preis ~ Kamin,   
 data = Houses)  
  
# t-Test  
t.test(Preis ~ Kamin,   
 data = Houses)

### Zwei Gruppen, kategorial (I/II)

# Säulendiagramm  
gf\_bar( ~ Kamin | Klimaanlage, data = Houses)



# Anteile  
prop(Kamin ~ Klimaanlage, data = Houses, success = "Ja")

## prop\_Ja.Ja prop\_Ja.Nein   
## 0.7559055 0.4647758

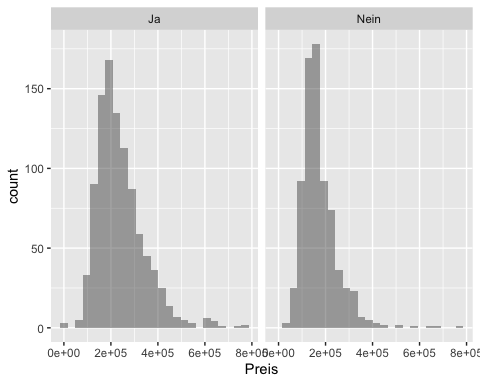
### Zwei Gruppen, kategorial (II/II)

# Anteilstest  
prop.test(Kamin ~ Klimaanlage, data = Houses, success = "Ja")

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: tally(Kamin ~ Klimaanlage)  
## X-squared = 137.85, df = 1, p-value < 2.2e-16  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## 0.2452697 0.3369896  
## sample estimates:  
## prop 1 prop 2   
## 0.7559055 0.4647758

### Zwei Gruppen, numerisch (I/II)

# Histogramm  
gf\_histogram( ~ Preis | Kamin, data = Houses)



# Mittelwerte  
mean(Preis ~ Kamin, data = Houses)

## Ja Nein   
## 239914.0 174653.4

### Zwei Gruppen, numerisch (II/II)

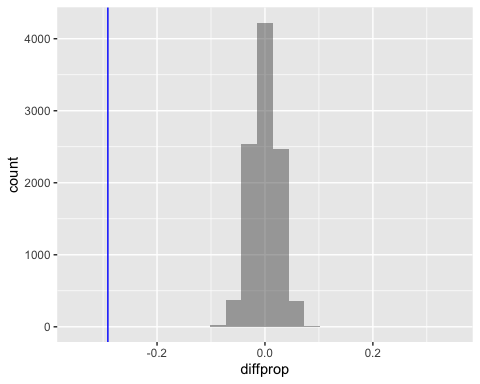
# t-Test  
t.test(Preis ~ Kamin, data = Houses)

##   
## Welch Two Sample t-test  
##   
## data: Preis by Kamin  
## t = 14.971, df = 1724.7, p-value < 2.2e-16  
## alternative hypothesis: true difference in means between group Ja and group Nein is not equal to 0  
## 95 percent confidence interval:  
## 56710.60 73810.61  
## sample estimates:  
## mean in group Ja mean in group Nein   
## 239914.0 174653.4

### Permutationstest kategorial

# Reproduzierbarkeit  
set.seed(2009)  
# Anteilsdifferenz in Stichprobe  
pdiff\_est <- diffprop(Kamin ~ Klimaanlage, success = "Ja", data = Houses)  
# Simuliere H\_0: Permutiere Klima  
Nullvtlg <- do(10000) \*   
 diffprop(Kamin ~ shuffle(Klimaanlage), success = "Ja", data = Houses)  
# Histogramm Nullverteilung  
gf\_histogram( ~ diffprop, data = Nullvtlg) %>%  
 gf\_vline(xintercept = ~ pdiff\_est, color = "blue") + xlim(-0.35, 0.35)

## Warning: Removed 2 rows containing missing values (geom\_bar).



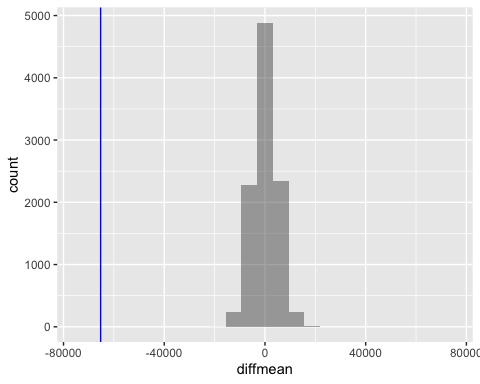
# p-Wert  
prop( ~(abs(diffprop)>=abs(pdiff\_est)), data = Nullvtlg )

## prop\_TRUE   
## 0

### Permutationstest numerisch

# Reproduzierbarkeit  
set.seed(2009)  
# Mittelwertdifferenz in Stichprobe  
meandiff\_est <- diffmean(Preis ~ Kamin, data = Houses)  
# Simuliere H\_0: Permutiere Klima  
Nullvtlg <- do(10000) \*   
 diffmean(Preis ~ shuffle(Kamin), data = Houses)  
# Histogramm Nullverteilung  
gf\_histogram( ~ diffmean, data = Nullvtlg) %>%  
 gf\_vline(xintercept = ~ meandiff\_est, color = "blue") +   
 xlim(-75000, 75000)

## Warning: Removed 2 rows containing missing values (geom\_bar).



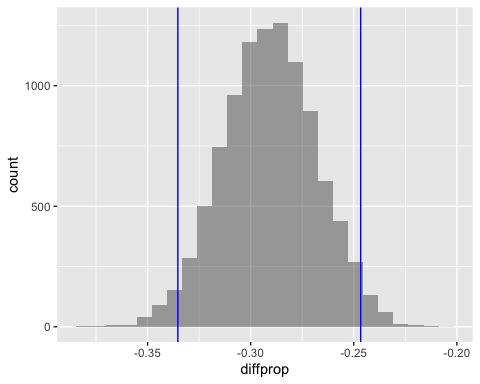
# p-Wert  
prop( ~(abs(diffmean)>=abs(meandiff\_est)), data = Nullvtlg)

## prop\_TRUE   
## 0

### Bootstrap kategorial

# Reproduzierbarkeit  
set.seed(2009)  
# Simuliere Stichprobenziehung  
Bootvtlg <- do(10000) \*   
 diffprop(Kamin ~ Klimaanlage, success = "Ja", data = resample(Houses))  
# 95% Konfidenzintervall  
ci <- quantile( ~ diffprop, probs = c(0.025, 0.975), data = Bootvtlg)  
# Histogramm  
gf\_histogram( ~ diffprop, data = Bootvtlg) %>%  
 gf\_vline(xintercept = ci, color = "blue")

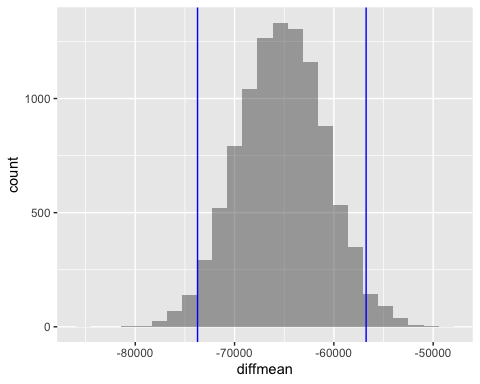
## Warning: geom\_vline(): Ignoring `mapping` because `xintercept` was provided.



### Bootstrap numerisch

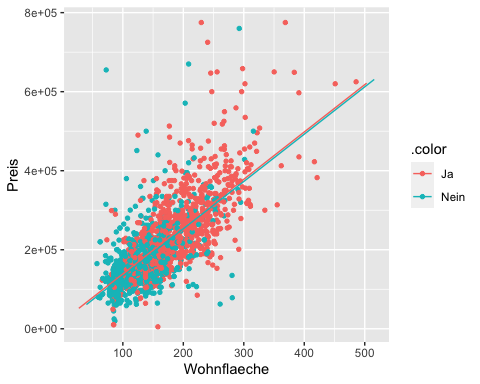
# Reproduzierbarkeit  
set.seed(2009)  
# Simuliere Stichprobenziehung  
Bootvtlg <- do(10000) \*   
 diffmean(Preis ~ Kamin, data = resample(Houses))  
# 95% Konfidenzintervall  
ci <- quantile( ~ diffmean, probs = c(0.025, 0.975), data = Bootvtlg)  
# Histogramm  
gf\_histogram( ~ diffmean, data = Bootvtlg) %>%  
 gf\_vline(xintercept = ci, color = "blue")

## Warning: geom\_vline(): Ignoring `mapping` because `xintercept` was provided.



### Lineares Modell (I/III)

modnum <- lm(Preis ~ Wohnflaeche + Kamin, data = Houses)  
plotModel(modnum)



### Lineares Modell (II/III)

summary(modnum)

##   
## Call:  
## lm(formula = Preis ~ Wohnflaeche + Kamin, data = Houses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -271421 -39935 -7887 28215 554651   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 19166.54 6286.94 3.049 0.00233 \*\*   
## Wohnflaeche 1197.15 31.94 37.476 < 2e-16 \*\*\*  
## KaminNein -5567.38 3716.95 -1.498 0.13436   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 69080 on 1725 degrees of freedom  
## Multiple R-squared: 0.5081, Adjusted R-squared: 0.5076   
## F-statistic: 891 on 2 and 1725 DF, p-value: < 2.2e-16

### Lineares Modell (III/III)

anova(modnum)

## Analysis of Variance Table  
##   
## Response: Preis  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Wohnflaeche 1 8.4934e+12 8.4934e+12 1779.8488 <2e-16 \*\*\*  
## Kamin 1 1.0706e+10 1.0706e+10 2.2435 0.1344   
## Residuals 1725 8.2317e+12 4.7720e+09   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

1. Befehlsübersicht [hier](https://github.com/luebby/Datenanalyse-mit-R/blob/master/Installation/R-mosaic-Kurzreferenz.pdf) [↑](#footnote-ref-20)
2. Siehe auch: ?mosaicData::SaratogaHouses [↑](#footnote-ref-25)