Wissenschaftliche Methodik: Poster-Wettbewerb

Lukas Küppers, Herndrick Kaiser und Vincent Schmalor

03.01.2020

# Einleitung

Dieses Dokument stellt sowohl die Durchführung, als auch die Protokollierung des Vorhersagewettbewerbs dar. Die Aufgabenstellung besteht darin, ein Modell aus einem Grunddatensatz zu entwickeln und dieses Modell auf einen Anwendungsdatensatz anzuwenden. Dabei wird aus mehreren Einflussvariablen eine abhängige Variable geschätzt. Der konkrete Anwendungsfall ist ein Auszug aus einem Face-to-Face Interview über das Konsum- und Schenkverhalten zu Weihnachten. Das Grunddatenset beinhaltet 400 Beobachtungen mit 17 erklärenden Variablen und die abhängige Variable des geschätzten Budgets, die es im Anwendungsdatenset zu modellieren gilt. In dem folgenden Kapitel wird diese Aufgabe bearbeitet, indem zunächst ein Überblick über die Daten in einer explorativen Datenanalyse gegeben wird. Anhand verschiedener statistischer Analysen und der Untersuchung der semantischen Zusammenhänge wird dann eine Vorgehensweise erarbeitet und beschrieben. Das Ergebnis ist eine csv-Datei, die das Anwendungsdatenset mit der modellierten, abhängigen Variable beinhaltet. So wird eine Prognose über diese Variable für jede Beobachtung gegeben, für die das tatsächliche Budget nicht vorliegt. Eine Zusammenfassung der Ergebnisse findet sich auch auf dem Plakat der genannten Autoren.

# Vorbereitung des Datensets

Nach Oliver Gansser empfiehlt es sich, vor der Analyse den Datensatz zu betrachten. Hierbei sind insbesondere die Datenstruktur von besonderer Relevanz. Außerdem fügt Oliver Gansser an, die Datenstruktur zu verändern. Aus diesem Grund wurde sich für die Vorbereitung des Datensets auf vier Punkte wie folgt festgelegt:

1. Die folgende Operation weist den Spalten des Datensets die korrekten sprechenden Bezeichnungen zu, um sie im späteren Verlauf besser identifizieren zu können

colnames(DST.raw) <- c("Beratung", "Angebote", "Bequemlichkeit", "Einkaufsatmosphaere", "Marken", "GPM", "Naehe", "Partner", "Eltern", "Verwandte", "Kommilitonen", "Kinder", "Freunde", "Arbeitskollegen", "GemNutz", "Alter", "Geschlecht", "Budget") #Vektor wird erstellt und auf die Spalten angewendet

1. Bevor die Daten verändert werden, wird an dieser Stelle die gesamte Stichprobe in mehrere Resamples gespeichert, um später etwa Bootstrap-Analysen durchführen zu können

set.seed(1996)  
DS.RS1 <- resample(DST.raw)  
set.seed(1993)  
DS.RS2 <- resample(DST.raw)  
set.seed(687654)  
DS.RS3 <- resample(DST.raw)  
set.seed(2020)

1. Das Datenset beinhaltet einige Ausreißer, die im Vorfeld der Analyse gefiltert werden sollten. Ausreißer sind sicher auch in der Population vorhanden. Da diese jedoch nicht modelliert werden können, werden sie für die Erstellung des Modells gefiltert.

DST.clean <- filter(DST.raw, Budget > 30) #Untere Elemente filtern. Realistische Werte sind ab 30€ zu erwarten  
OL.DS.raw <- boxplot.stats(DST.clean$Budget)$out #Boxplotergebnisse verwenden, um Ausreißer zu speichern  
paste("Ausreißer: ", paste(OL.DS.raw, collapse=", ")) #Ausreißerwerte ausgeben

## [1] "Ausreißer: 1500, 2500, 1500, 1500, 3100, 3500, 1500, 1300, 1400, 2500, 1500, 1500, 1500, 1500, 2100, 1500, 2500, 2000, 1285, 1500"

DST.clean <- DST.clean %>% filter(!(Budget %in% OL.DS.raw)) #Filter anwenden

1. Im Folgenden werden zwei abgeleitete Kennzahlen eingeführt. Die Kennzahlen leiten sich aus den Personen ab, die beschenkt werden, und welche Ansprüche die Schenkenden an die Geschenke haben.

-Anzahl der Geschenke: Die Vermutung besteht, dass ein Zusammenhang zwischen der Gesamtanzahl an beschenkten Personen und dem geplanten Budget besteht. Die Summe der beschenkten Personen wird daher im Folgenden als Anzahl der Geschenke verwendet.

-Wertschätzung: Die Vermutung besteht, dass ein Zusammenhang zwischen der Summe der als wichtig bezeichneten Geschenk- und Dienstleistungseigenschaften und dem geplanten Budget besteht. Die Summe der Wertschätzungsgrade wird daher im Folgenden als Wertschätzung verwendet.

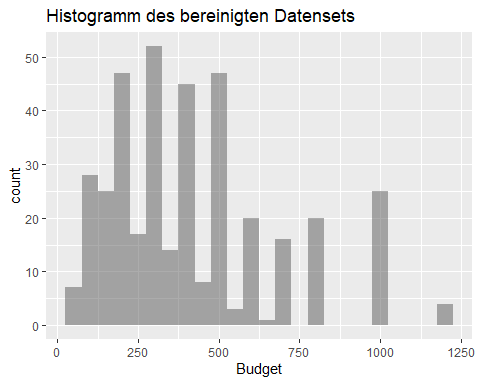
DST.raw.extra <- mutate(DST.raw, GesAnz = Partner + Eltern + Verwandte + Kommilitonen + Kinder + Freunde + Arbeitskollegen + GemNutz) #Geschenkanzahl hinzufügen  
DST.raw.extra <- mutate(DST.raw.extra, Wertsch = Beratung + Angebote + Bequemlichkeit + Einkaufsatmosphaere + Marken + GPM + Naehe) #Wertschätzung hinzufügen  
DST.extra <- mutate(DST.clean, GesAnz = Partner + Eltern + Verwandte + Kommilitonen + Kinder + Freunde + Arbeitskollegen + GemNutz)  
DST.extra <- mutate(DST.extra, Wertsch = Beratung + Angebote + Bequemlichkeit + Einkaufsatmosphaere + Marken + GPM + Naehe)

# Explorative Datenanalyse

## Verteilungen untersuchen

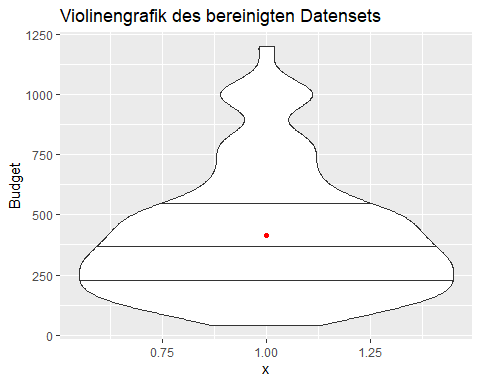
Im Folgenden werden die Daten verschieden dargestellt, um einen Überblick über die Verteilungen zu bekommen. Hierbei zeigen ein Histogramm und eine Violinengrafik die bereits bereinigten Datensätze. Zur besseren Veranschaulichung zeigen die abschließenden Kastengrafiken sowohl die unbereinigten als auch die bereinigten Datensätze.

### Histogramm



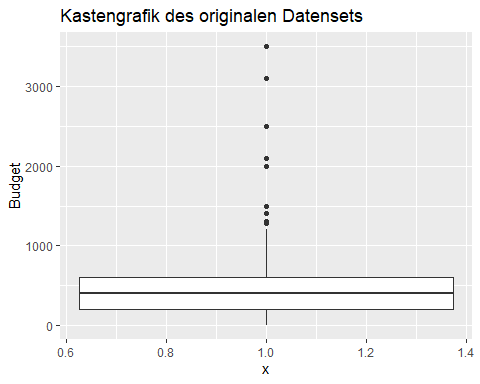
Die Verteilung mit vielen Spitzen lässt sich darauf zurückführen, dass bei denb Budgetangaben lediglich grobe Schätzungen abgegeben werden und diese in den meisten Fällen auf Hunderter gerundet werden.

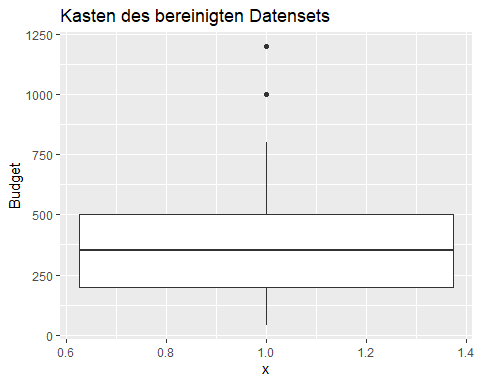
### Violinengrafik



Die Violinengrafik zeigt, dass sich der Großteil der Daten im Bereich zwischen 250€ und 600€ befindet. Wenige Werte über dieser Grenze verschieben das arithmetische Mittel über den Median.

### Kastengrafik





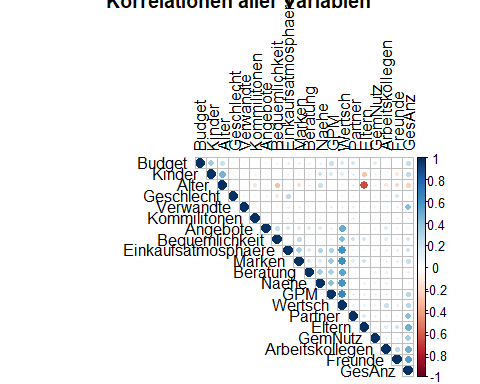
In der ersten Kastengrafik wird das ubereinigte Spektrum der Daten visualisiert. Die zweite Kastengrafik zeigt im Kontrast dazu die im oben beschriebenen Schritt bereinigten Daten.

## Zusammenhänge finden

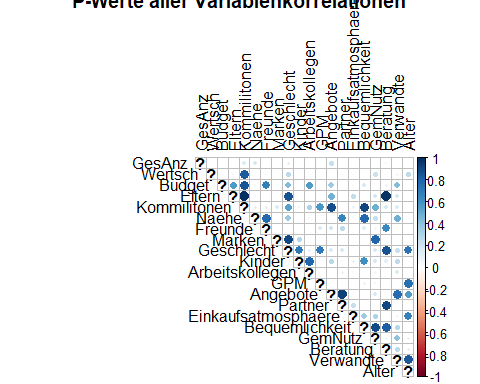
Nachdem die Grunddatenmenge visualisiert wurde, müssen für die Erstellung eines Modells die Zusammenhänge zwischen den Variablen untersucht werden. Insbesondere die Relationen zwischen allen Variablen und dem Budget sind hierbei von Relevanz. Mit der Funktion “rcorr” wird eine Korrelationsmatrix erstellt. Anhand dieser Korrelationsmatrix lassen sich Relationen und P-Werte ablesen. Diese Werte bieten eine datenbasierte Möglichkeit, die statistische Relevanz der Korrelation zu bestimmen. So können zufällig starke Korrelationen von der weiteren Betrachtung ausgeschlossen werden.

DST.mat <- as.matrix(data.frame(DST.num)) #Konvertierung in verwendbares Format  
CMT <- rcorr(DST.mat, type = "pearson") #Erstellung der Korrelationsmatrix

corrplot(CMT$r, type = "upper", order = "hclust", tl.col = "black", sig.level = 0.01, title = "Korrelationen aller Variablen") #Visualisierung der Korrelationen aller Variablen untereinander

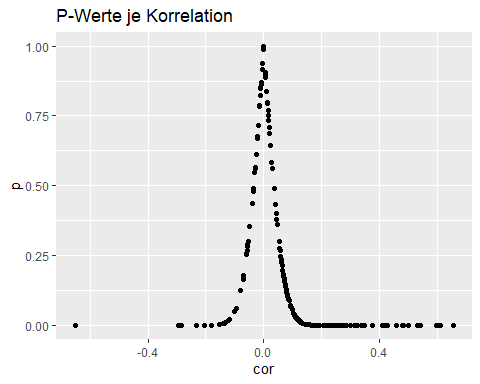


corrplot(CMT$P, type = "upper", order = "hclust", tl.col = "black", sig.level = 0.01, title = "P-Werte aller Variablenkorrelationen") #Visualisierung der P-Werte aller Korrelationen zwischen allen Variablen



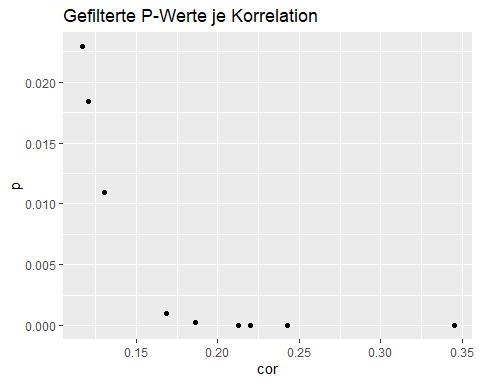
Erwartungsgemäß bestehen starke Korrelationen zwischen Wertschätzung und deren Einflussvariablen sowie der Geschenkanzahl und deren Einflussfaktoren. Eine starke negative Korrelation besteht auch zwischen dem Alter und der Frage, ob Eltern beschenkt werden. Diese Erkenntnisse sind sicher einerseits interessant, andererseits auch erwartbar, aber nicht relevant in der Modellierung des Budgets. Da es keine starken Ausprägungen in den Relationen zwischen Budget und anderen Variablen gibt, wird in der nächsten Grafik der P-Wert je Korrelation dargestellt.

CMT.tbl <- flattenCorrMatrix(CMT$r, CMT$P) #Tabellen joinen, um eine Tabelle zu erhalten  
gf\_point(p ~ cor, data = CMT.tbl, title = "P-Werte je Korrelation") #Visualisierung



Wie vermutet weisen viele Korrelationen einen viel zu hohen P-Wert aus, als dass sie in der Modellierung des Budgets verwendet werden könnten. Eine Filterung auf einen angemessenen P-Wert ist daher unerlässlich.

CMT.tbl.f <- filter(CMT.tbl, p < 0.03 & (row == "Budget" | column == "Budget"))  
gf\_point(p ~ cor, data = CMT.tbl.f, title = "Gefilterte P-Werte je Korrelation")



CMT.tbl.f

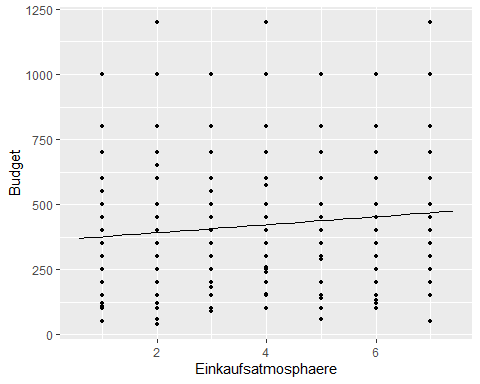
## row column cor p  
## 1 Einkaufsatmosphaere Budget 0.1209947 0.018453103487469535  
## 2 Marken Budget 0.1168220 0.022932215428778102  
## 3 GPM Budget 0.2128488 0.000029407348239641  
## 4 Partner Budget 0.1860446 0.000270757664015076  
## 5 Kinder Budget 0.3448485 0.000000000005047518  
## 6 GemNutz Budget 0.1305671 0.010947326592058548  
## 7 Alter Budget 0.2197324 0.000015854409149529  
## 8 Budget GesAnz 0.2425775 0.000001766615588927  
## 9 Budget Wertsch 0.1682677 0.001007301360662405

Die Grafik zeigt die relativen Positionierungen von P-Werten und Korrelationsstärken. Sie korrespondieren mit den Werten der Tabelle. Nun, da die relevanten Einflussfaktoren isoliert wurden, werden ausschließlich diese Variablen weiterverwendet. Aus der erstellten Grafik und der abgeleiteten Tabelle weisen die Einkaufsatmosphaere (≈0.0185), die Marken (≈0.2293) und GemNutz (≈0.0109) die höchsten P-Werte auf. Diese Variablen sollten in der Analyse kritisch betrachtet und gegebenenfalls nicht verwendet werden.

## Lineare Regressionen der relevanten Einflussvariablen

Im vorigen Kapitel wurden relevante Korrelationen ausgewählt. In diesem Kapitel werden drei Relationen in linearen modellen beispielhaft visualisiert. Für die Skalenniveaus der kategoriellen (Likertskala), der binären und der numerischen Variablen wird je ein Modell dargestellt.

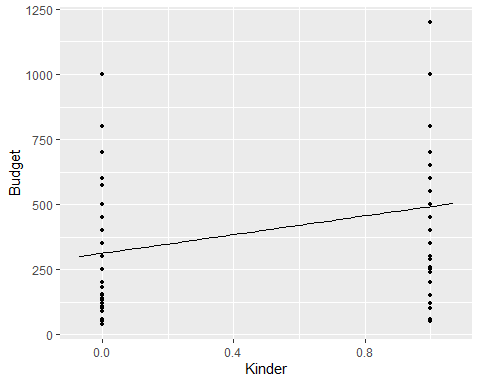
#Einkaufsatmosphäre  
lm.budget.Einkaufsatmosphaere <- lm(Budget ~ Einkaufsatmosphaere, data = DST.num)  
plotModel(lm.budget.Einkaufsatmosphaere, title = "Weihnachtsbudget in Relation zur Variable Einkaufsatmosphaere")



summary(lm.budget.Einkaufsatmosphaere)

##   
## Call:  
## lm(formula = Budget ~ Einkaufsatmosphaere, data = DST.num)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -418.42 -191.75 -57.08 123.58 808.25   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 361.083 26.529 13.611 <0.0000000000000002 \*\*\*  
## Einkaufsatmosphaere 15.334 6.479 2.367 0.0185 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 257 on 377 degrees of freedom  
## Multiple R-squared: 0.01464, Adjusted R-squared: 0.01203   
## F-statistic: 5.601 on 1 and 377 DF, p-value: 0.01845

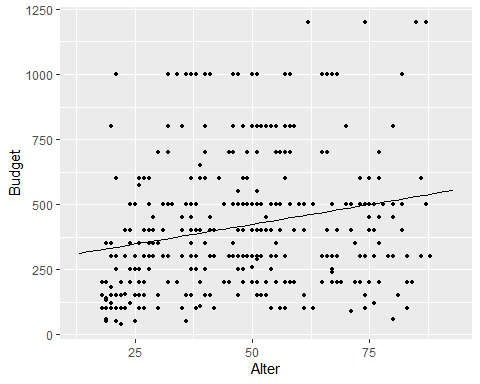
#Kinder  
lm.budget.Kinder <- lm(Budget ~ Kinder, data = DST.num)  
plotModel(lm.budget.Kinder, title = "Weihnachtsbudget in Relation zur Variable Kinder")



summary(lm.budget.Kinder)

##   
## Call:  
## lm(formula = Budget ~ Kinder, data = DST.num)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -442.49 -177.47 -42.49 107.51 707.51   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 312.47 19.10 16.363 < 0.0000000000000002 \*\*\*  
## Kinder 180.02 25.24 7.133 0.00000000000505 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 243 on 377 degrees of freedom  
## Multiple R-squared: 0.1189, Adjusted R-squared: 0.1166   
## F-statistic: 50.88 on 1 and 377 DF, p-value: 0.000000000005048

#Alter  
lm.budget.alter <- lm(Budget ~ Alter, data = DST.num)  
plotModel(lm.budget.alter, title = "Weihnachtsbudget in Relation zur Variable Alter")



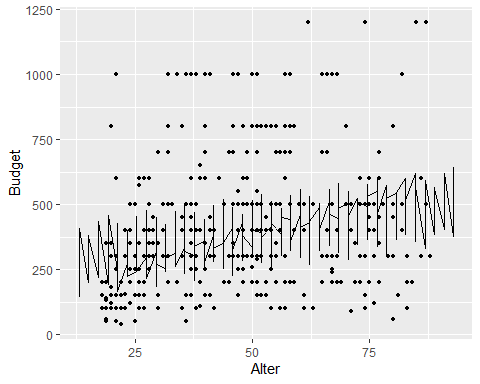
summary(lm.budget.alter)

##   
## Call:  
## lm(formula = Budget ~ Alter, data = DST.num)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -455.55 -192.46 -48.05 112.36 739.27   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 271.9164 35.3115 7.701 0.000000000000121 \*\*\*  
## Alter 3.0454 0.6964 4.373 0.000015854409150 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 252.6 on 377 degrees of freedom  
## Multiple R-squared: 0.04828, Adjusted R-squared: 0.04576   
## F-statistic: 19.13 on 1 and 377 DF, p-value: 0.00001585

# Erstellung des Modells

In den vorigen Kapiteln wurde das Datenset bearbeitet, veranschaulicht und es wurden die relevanten Einflussfaktoren auf das Budget der befragten Personen identifiziert. Um das Budget zu modellieren wird an dieser Stelle eine multiple Regression mit den identifizierten Einflussfaktoren vorgenommen. Anhand dieser Regressionsanalyse ist es möglich, ein unvollständiges Datenset zu extrapolieren. In das Modell fließen ausschließlich die relevantesten Variablen ein, um ein Overfitting zu vermeiden und gegebenenfalls zufällige Einflüsse zu filtern. Da selbst unter den ausgewählten Einflüssen noch immer hohe P-Werte bestehen, werden in der Regression nicht alle Einflüsse berücksichtigt. Das Budget wird anhand des Alters, der Güte der Produkte und Marken, der Anzahl an beschenkten Personen und der Tatsache, ob Kinder und Partner beschenkt werden, erstellt. Diese Variablen modellieren daher in diesem Fall das Budget.

lm.Gesamt <- lm(  
 Budget ~ Alter + GPM + GesAnz + Partner + Kinder , data = DST.num  
 ) #Modell erstellen  
plotModel(lm.Gesamt) #Modell visualisieren



summary(lm.Gesamt) #Modell in Zahlen ausdrücken

##   
## Call:  
## lm(formula = Budget ~ Alter + GPM + GesAnz + Partner + Kinder,   
## data = DST.num)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -478.77 -155.08 -44.62 109.26 766.27   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.882 58.109 0.239 0.81132   
## Alter 2.505 0.798 3.139 0.00183 \*\*  
## GPM 16.965 6.802 2.494 0.01305 \*   
## GesAnz 32.608 11.874 2.746 0.00632 \*\*  
## Partner 63.387 30.170 2.101 0.03631 \*   
## Kinder 98.857 30.237 3.269 0.00118 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 234.4 on 373 degrees of freedom  
## Multiple R-squared: 0.1895, Adjusted R-squared: 0.1787   
## F-statistic: 17.44 on 5 and 373 DF, p-value: 0.000000000000001579

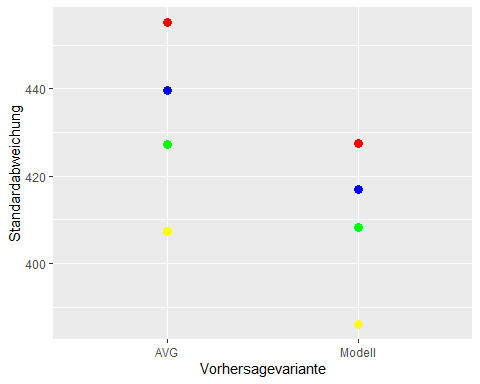
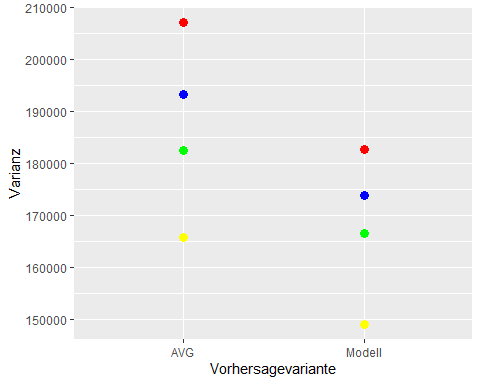
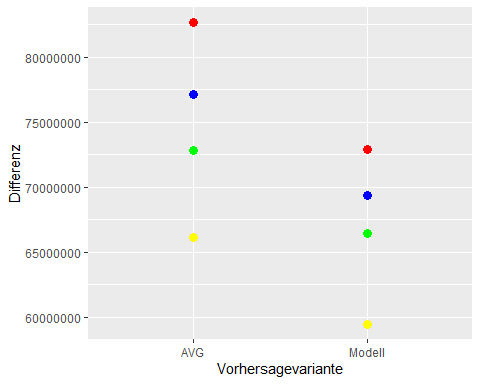
Das Ergebnis zeigt, dass auch in Kombination der Variablen in einem Modell weiterhin hohe Signifikanzniveaus bestehen. Das Modell erklärt knapp 18% der Varianz. Einflüsse, wie etwa der sozioökonomische Status der Personen würden die Güte des Modells vermutlich steigern.

# Validierung des Modells

Um die Güte des Modells zu prüfen, wird das Modell auf die resampleten Daten angewandt und mit dem Durchschnitt verglichen. Die Methodik ist hier zwei Mal gelistet und wird vier Mal angewandt. Da die Budgets in den Datensets tatsächlich vorhanden sind, kann der geschätzte Wert mit dem tatsächlichen Wert verglichen werden. Dazu werden die quadrierten Differenzen summiert. Aus dieser Kennzahl werden die Varianz und die Standardabweichung gebildet. Der direkte Vergleich des Nullmodells mit dem erarbeiteten Modell soll dann einen Anhaltspunkt dafür geben, ob das erarbeitete Modell bessere Schätzungen abgibt, als das Nullmopdell.

#Mean erstellen  
DST.AVG <- DST.raw %>% mutate(Budget2 = mean(DST.raw$Budget)) #Durchschnitt berechnen  
DST.AVG <- DST.AVG %>% mutate(SQDiff = (Budget - Budget2)^2) #Differenzen berechnen  
avg.diff <- sum(DST.AVG$SQDiff) #Differenzen summieren  
var.DST.AVG <- (1/399) \* avg.diff #Varianz berechnen  
sd.DST.AVG <- sqrt(var.DST.AVG) #Standardabweichung berechnen  
#paste("Trainingsdaten: Die quadrierte Differenz der tatsächlichen Werte zum Durchschnitt beträgt ", avg.diff) #Differenz ausgeben  
#Modell anwenden  
DST.erglm <- predict.lm(lm.Gesamt, newdata = DST.raw.extra, interval="prediction") #Modell anwenden  
DST.erglm <- head(matrix(DST.erglm), n=400) #Ergebnis kürzen  
DST.pred <- DST.raw %>% mutate(Budget2 = DST.erglm) #Ergebnis anfügen  
DST.pred <- DST.pred %>% mutate(SQDiff = (Budget - Budget2)^2) #Differenzen berechnen  
diff.DST.pred <- sum(DST.pred$SQDiff) #Differenzen summieren  
var.DST.pred <- (1/399)\*diff.DST.pred #Varianz berechnen  
sd.DST.pred <- sqrt(var.DST.pred) #Standardabweichung berechnen  
#paste("Trainingsdaten: Die quadrierte Differenz der tatsächlichen Werte zum Modell beträgt ", diff.DST.pred) #Differenz ausgeben

#Mean erstellen  
DS.RS1.AVG <- DS.RS1 %>% mutate(Budget2 = mean(DST.raw$Budget)) #Durchschnitt berechnen  
DS.RS1.AVG <- DS.RS1.AVG %>% mutate(SQDiff = (Budget - Budget2)^2) #Differenzen berechnen  
diff.DS.RS1.AVG <- sum(DS.RS1.AVG$SQDiff) #Differenzen summieren  
var.DS.RS1.AVG <- (1/399) \* diff.DS.RS1.AVG #Varianz berechnen  
sd.DS.RS1.AVG <- sqrt(var.DS.RS1.AVG) #Standardabweichung berechnen  
#paste("Resample 1: Die quadrierte Differenz der tatsächlichen Werte zum Durchschnitt beträgt ", diff.DS.RS1.AVG) #Differenz ausgeben  
DS.RS1.extra <- mutate(DS.RS1, GesAnz = Partner + Eltern + Verwandte + Kommilitonen + Kinder + Freunde + Arbeitskollegen + GemNutz) #Geschenkanzahl hinzufügen  
DS.RS1.extra <- mutate(DS.RS1.extra, Wertsch = Beratung + Angebote + Bequemlichkeit + Einkaufsatmosphaere + Marken + GPM + Naehe) #Wertschätzung hinzufügen  
DS.RS1.erglm <- predict.lm(lm.Gesamt, newdata = DS.RS1.extra, interval="prediction") #Modell anwenden  
DS.RS1.erglm <- head(matrix(DS.RS1.erglm), n=400) #Ergebnis kürzen  
DS.RS1.pred <- DS.RS1.extra %>% mutate(Budget2 = DS.RS1.erglm) #Ergebnisanfügen  
DS.RS1.pred <- DS.RS1.pred %>% mutate(SQDiff = (Budget - Budget2)^2) #Differenzen berechnen  
diff.DS.RS1.pred <- sum(DS.RS1.pred$SQDiff) #Differenzen summieren  
var.DS.RS1.pred <- (1/399) \* diff.DS.RS1.pred #Varianz berechnen  
sd.DS.RS1.pred <- sqrt(var.DS.RS1.pred) #Standardabweichung berechnen  
#paste("Resample 1: Die quadrierte Differenz der tatsächlichen Werte zum Durchschnitt beträgt ", diff.DS.RS1.pred)



Die Grafiken zeigen, dass die Differenz, die Varianz und die Standardabweichung des erarbeiteten Modells in allen Fällen unter denen des Nullmodells liegen. Damit besteht die Vermutung, dass das erarbeitete Modell eine bessere Vorhersage ausgeben wird, als das Nullmodell.

# Anwendung des Modells

Wie in der Einleitung beschrieben soll das Modell auf ein Datenset angewendet werden, das keine Budgetwerte besitzt. Anhand des Modells wird an dieser Stelle für jede Beobachtung das geschätzte Budget geschätzt. Das Ergebnis ist ein Vektor mit geschätzten Budgets, der den entsprechenden Beobachtungen zugeordnet wird.

DSA.raw <- read.csv2("sources/Anwendungsdaten.csv")  
colnames(DSA.raw) <- c("Beratung", "Angebote", "Bequemlichkeit", "Einkaufsatmosphaere", "Marken", "GPM", "Naehe", "Partner", "Eltern", "Verwandte", "Kommilitonen", "Kinder", "Freunde", "Arbeitskollegen", "GemNutz", "Alter", "Geschlecht") #Vektor wird erstellt und auf die Spalten angewendet  
DSA.extra <- mutate(DSA.raw, GesAnz = Partner + Eltern + Verwandte + Kommilitonen + Kinder + Freunde + Arbeitskollegen + GemNutz)  
DSA.extra <- mutate(DSA.extra, Wertsch = Beratung + Angebote + Bequemlichkeit + Einkaufsatmosphaere + Marken + GPM + Naehe)  
DSA.erglm <- predict.lm(lm.Gesamt, newdata = DSA.extra, interval="prediction") #Modell anwenden  
DSA.erglm <- head(matrix(DSA.erglm), n= 600)  
DSA.fin <- mutate(DSA.extra, Budget = DSA.erglm)  
write.csv2(DSA.erglm, "Prognose\_Vincent\_Schmalor\_Hendrick\_Kaiser\_Lukas\_Kueppers.csv")

# Untersuchung der Ergebnisse

In diesem Kapitel soll untersucht werden, ob die ermittelten Werte plausibel sind oder unerwartete Werte auftreten. Ziel ist keine eingehende Untersuchung der Einzelfälle, sondern eine Validierung des Gesamtbildes. Unerwartete Werte wären an dieser Stelle etwa negative Budgets oder Budgets über 1000€. Eine etwa gleichmäßige Verteilung ist ebenfalls anzunehmen.

## Übersicht

Die folgenden Kennzahlen zeigen, dass es keine extremen Ausreißer gibt und Median und Durchschnitt in einem nach den Verteilungsuntersuchen in Kapitel “Explorative Datenanalyse” erwartbaren Bereich liegen. Die Grafik bekräftigt diesen Eindruck.

## min Q1 median Q3 max mean sd n missing  
## 116.0586 343.5556 428.3234 508.3975 650.802 421.1625 107.2871 600 0

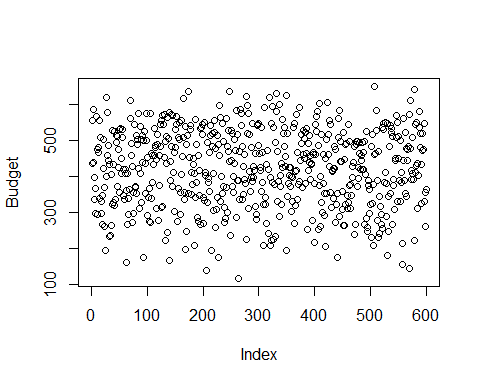
## Warning in plot.window(...): "title" ist kein Grafikparameter

## Warning in plot.xy(xy, type, ...): "title" ist kein Grafikparameter

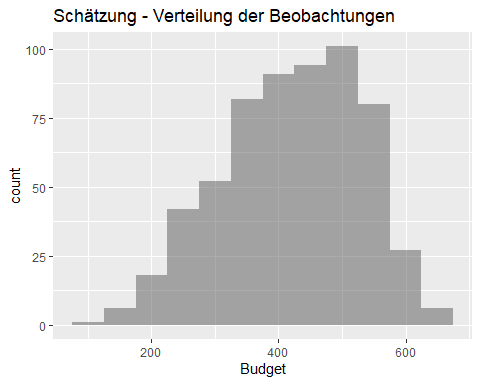
## Warning in axis(side = side, at = at, labels = labels, ...): "title" ist  
## kein Grafikparameter  
  
## Warning in axis(side = side, at = at, labels = labels, ...): "title" ist  
## kein Grafikparameter

## Warning in box(...): "title" ist kein Grafikparameter

## Warning in title(...): "title" ist kein Grafikparameter

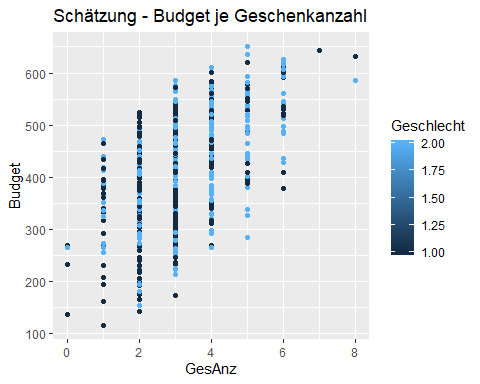
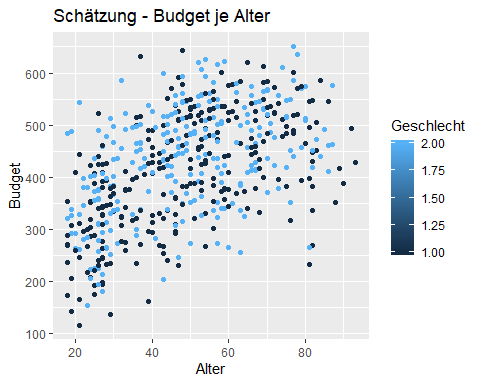


## Histogramm



Das Histogramm zeigt eine Verteilung der Budgets nach Höhe ebendessen.

## Punktwolken



Die untersuchte Korrelationen von Budget und Alter und der Geschenkanzahl lassen sich auch in den geschätzten Werten finden.

# Fazit

Die Untersuchung der Stichprobe und die anschließende Modellierung haben einige interessante Informationen aufgedeckt. Die Forschungsfrage etwa, ob es eine statistisch relevante Korrelation zwischen dem Budget und dem Geschlecht gibt, lässt sich klar verneinen. Starke Einflussfaktoren mit hoher Relevanz sind allerdings, wie viele Personen beschenkt werden und ob Kinder beschenkt werden.

Die explorative Datenanalyse hat zunächst wertvolle Informationen über Verteilungen und Schwerpunkte der Daten gegeben. Diese Verteilung lässt sich später auch in den geschätzten Werten wiederfinden. Ein Überblick über die Korrelationen und deren Relevanz hat zu einer engen Auswahl an relevanten Variablen geführt. Die auf diesem Weg ermittelten Variablen sind in das Modell eingeflossen und wurden damit bei der Schätzung der unbekannten Werte im Zieldatenset verwendet.

Mit Hilfe der resampleten Stichproben konnte ein Hinweis auf die Validitat des Modells gegeben werden. Eine Bestätigung darüber, ob das gewählte Modell auch die unbekannten Daten gut vorhersagt, ist das allerdings nicht. In den vier Durchgängen wurde mit dem entwickelten Modell immer eine geringere Standardabweichung erzielt, als es das Nullmodell tat. Die Untersuchung der geschätzten Werte hat ebenfalls ergeben, dass das Modell korrekt angewendet wurde. Die Zusammenhänge von Alter, Geschenkanzahl und Budget lassen sich in den geschätzten Werten wiederfinden. Auch die oberen und unteren Werte befinden sich in etwa auf dem Niveau, wie es die Budgets der Trainingsdaten taten. Eine gewisse Aussagekraft des Modells lässt sich auch aus dem Wert des bereinigten R-Quadrats der linearen Regression von knapp 18% und der hohen statistischen Relevanz, die durch den sehr geringen P-Wert ausgedrückt wird, ableiten.

Zusammenfassend kann man festhalten, dass das festgelegte Ziel der Datenanalyse, -modellierung und -validierung erreicht wurde. Das Modell schätzt die Werte in einem erwartbaren Bereich unter der Berücksichtigung rationaler Zusammenhänge und weist bessere Werte in der Validierung auf als das arithmetische Mittel. Allerdings liegt die Standardabweichung bei etwa 400€. Auch das R-Quadrat von knapp 19% belegt, dass ein Großteil der Varianz nicht erklärt werden konnte. Für eine bessere Modellierung müssten weitere Variablen wie etwa das Einkommensniveau vorliegen. Anhand der vorliegenden Daten wurde daher ein Modell entwickelt, das die Budgets mit gesteigerter Genauigkeit schätzt, ohne dabei zu sehr auf die Trainingsdaten zugeschnitten zu sein.