# Projektarbeit zu Algorithmik & Statistik für Data Science 2 Lab

## Black Friday - eine statistische Analyse der Verkaufszahlen

Christoph Rabensteiner: 1810837995

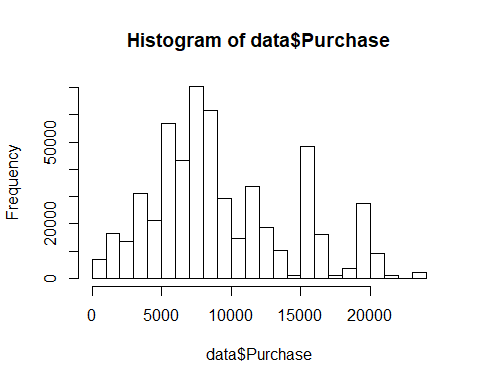
Valentin Muhr: 1810837102

Magdalena Breu: 1810837995

Jochen Paul Hollich: 1810837475

## Beschreibung des Datensatzes.

Der Datensatz Black Friday enthält Daten über den Warenkorb von Kunden bei einem Onlineshop. Im Datensatz befinden sich 537577 Beobachtungen für 12 Variablen, u.a. Geschlecht, Alter, Beruf, Beziehungsstatus und Kaufkraft. Die Variablen Age, Occupation, City\_Category, Stay\_In\_Current\_City\_Years Product\_Categories sind als Faktoren codiert. Die genauen Definitionen der einzelnen Faktoren (z.B. City\_Category A,B &C) sind nicht genau definiert.



## Hintergrundinformationen zum Datensatz

Der Datensatz wurde bereitgestellt durch einen Hackathon der Firma Analytics Vidhya (<https://datahack.analyticsvidhya.com/contest/black-friday/>) und heruntergeladen über Kaggle.com (<https://www.kaggle.com/mehdidag/black-friday>).

## Die statistische Lernaufgabe

Mit diesem Datensatz soll eine Customer Segmentierung vorgenommen werden, um so den einzelnen Kundengruppen gezielt z.B. Werbung einschalten zu können oder als Input für einen Report an die Geschäftsleitung. Unter anderem sind die folgenden Analysen vorgesehen:

- Segmentierung der Customerbase mittles Clustering  
 - Klassifizierung z.B. nach Stadt, Berufsgruppe etc.   
 - Regression: Vorhersage des Sale-Volumen für die jeweiligen Customer Segmente

## Gründe warum die statistische Lernaufgabe auf diesen Datensatz angewendet wird

Ausser für den Abschluss dieses Projekts, haben wir uns für den Black Friday Datensatz entschieden, da alle von uns einen wirtschaftlichen Hintergrund haben (Studium oder Beruf) und uns erhoffen, die Modelle, die wir erstellen auch verstehen und interpretieren koennen.

Der Datensatz ist ein Beispiel für eine Transaktionsliste für einen Einzelhandelsshop. Der Shop moechte das Verhalten seiner Kunden zu ihren Produkten besser verstehen. Wir sind der Meinung, dass dies ein realistisches Szenario ist, welches wir auch spaeter in unserem beruflichen Werdegang gut gebrauchen koennen. So kann beispielsweise ein

## Nachweis, dass die Daten in R geladen werden können

summary(data)

## User\_ID Product\_ID Gender Age   
## Min. :1000001 P00265242: 1858 F:132197 0-17 : 14707   
## 1st Qu.:1001495 P00110742: 1591 M:405380 18-25: 97634   
## Median :1003031 P00025442: 1586 26-35:214690   
## Mean :1002992 P00112142: 1539 36-45:107499   
## 3rd Qu.:1004417 P00057642: 1430 46-50: 44526   
## Max. :1006040 P00184942: 1424 51-55: 37618   
## (Other) :528149 55+ : 20903   
## Occupation City\_Category Stay\_In\_Current\_City\_Years  
## Min. : 0.000 A:144638 0 : 72725   
## 1st Qu.: 2.000 B:226493 1 :189192   
## Median : 7.000 C:166446 2 : 99459   
## Mean : 8.083 3 : 93312   
## 3rd Qu.:14.000 4+: 82889   
## Max. :20.000   
##   
## Marital\_Status Product\_Category\_1 Product\_Category\_2 Product\_Category\_3  
## Min. :0.0000 Min. : 1.000 Min. : 2.00 Min. : 3.0   
## 1st Qu.:0.0000 1st Qu.: 1.000 1st Qu.: 5.00 1st Qu.: 9.0   
## Median :0.0000 Median : 5.000 Median : 9.00 Median :14.0   
## Mean :0.4088 Mean : 5.296 Mean : 9.84 Mean :12.7   
## 3rd Qu.:1.0000 3rd Qu.: 8.000 3rd Qu.:15.00 3rd Qu.:16.0   
## Max. :1.0000 Max. :18.000 Max. :18.00 Max. :18.0   
## NA's :166986 NA's :373299   
## Purchase   
## Min. : 185   
## 1st Qu.: 5866   
## Median : 8062   
## Mean : 9334   
## 3rd Qu.:12073   
## Max. :23961   
##

data[1:4,]

## User\_ID Product\_ID Gender Age Occupation City\_Category  
## 1 1000001 P00069042 F 0-17 10 A  
## 2 1000001 P00248942 F 0-17 10 A  
## 3 1000001 P00087842 F 0-17 10 A  
## 4 1000001 P00085442 F 0-17 10 A  
## Stay\_In\_Current\_City\_Years Marital\_Status Product\_Category\_1  
## 1 2 0 3  
## 2 2 0 1  
## 3 2 0 12  
## 4 2 0 12  
## Product\_Category\_2 Product\_Category\_3 Purchase  
## 1 NA NA 8370  
## 2 6 14 15200  
## 3 NA NA 1422  
## 4 14 NA 1057

## Einfache Modellierung der Daten

Als einfaches Beispiel für die mögliche Modellierung der Black Friday Daten wird eine einfache lineare Regression auf die ersten 200 Datensätze gefittet. Die ersten 10 Predictions sind ausgegeben.

library(caret)  
training <- data[1:200,]  
training$Product\_ID <- NULL  
  
model\_test <- train(  
 Purchase ~ .,   
 data = training,   
 method = "lm",   
 preProc = c("center", "scale"),  
 na.action = na.pass   
)  
  
pred\_test <- predict(model\_test, training)  
pred\_test[1:10]

## 2 7 14 15 17 19 20   
## 15427.779 16407.871 12212.433 7838.821 12169.861 15129.572 16484.783   
## 25 29 30   
## 16019.037 4331.582 14530.475