Projektbericht

Ferdinand Bubeck und Johannes Bubeck

19.01.2021

## Data Exploration

#import libraries  
library(readr)  
library(dplyr)  
library(ggplot2)  
library(tidyverse)  
  
#import CSVs  
datTest <- read\_csv("aug\_test.csv")  
  
datTrain <- read\_csv("aug\_train.csv")  
  
#join of Train & Test Data to have full Dataset for Exploration  
datFull <- full\_join(datTest, datTrain)   
  
#first Impressions of Data  
head(datFull)

## # A tibble: 6 x 14  
## enrollee\_id city city\_developmen~ gender relevent\_experi~ enrolled\_univer~  
## <dbl> <chr> <dbl> <chr> <chr> <chr>   
## 1 32403 city~ 0.827 Male Has relevent ex~ Full time course  
## 2 9858 city~ 0.92 Female Has relevent ex~ no\_enrollment   
## 3 31806 city~ 0.624 Male No relevent exp~ no\_enrollment   
## 4 27385 city~ 0.827 Male Has relevent ex~ no\_enrollment   
## 5 27724 city~ 0.92 Male Has relevent ex~ no\_enrollment   
## 6 217 city~ 0.899 Male No relevent exp~ Part time course  
## # ... with 8 more variables: education\_level <chr>, major\_discipline <chr>,  
## # experience <chr>, company\_size <chr>, company\_type <chr>,  
## # last\_new\_job <chr>, training\_hours <dbl>, target <dbl>

tail(datFull)

## # A tibble: 6 x 14  
## enrollee\_id city city\_developmen~ gender relevent\_experi~ enrolled\_univer~  
## <dbl> <chr> <dbl> <chr> <chr> <chr>   
## 1 29754 city~ 0.92 Female Has relevent ex~ no\_enrollment   
## 2 7386 city~ 0.878 Male No relevent exp~ no\_enrollment   
## 3 31398 city~ 0.92 Male Has relevent ex~ no\_enrollment   
## 4 24576 city~ 0.92 Male Has relevent ex~ no\_enrollment   
## 5 5756 city~ 0.802 Male Has relevent ex~ no\_enrollment   
## 6 23834 city~ 0.855 <NA> No relevent exp~ no\_enrollment   
## # ... with 8 more variables: education\_level <chr>, major\_discipline <chr>,  
## # experience <chr>, company\_size <chr>, company\_type <chr>,  
## # last\_new\_job <chr>, training\_hours <dbl>, target <dbl>

summary(datFull)

## enrollee\_id city city\_development\_index gender   
## Min. : 1 Length:21287 Min. :0.4480 Length:21287   
## 1st Qu.: 8554 Class :character 1st Qu.:0.7390 Class :character   
## Median :16967 Mode :character Median :0.9030 Mode :character   
## Mean :16874 Mean :0.8285   
## 3rd Qu.:25162 3rd Qu.:0.9200   
## Max. :33380 Max. :0.9490   
##   
## relevent\_experience enrolled\_university education\_level major\_discipline   
## Length:21287 Length:21287 Length:21287 Length:21287   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## experience company\_size company\_type last\_new\_job   
## Length:21287 Length:21287 Length:21287 Length:21287   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## training\_hours target   
## Min. : 1.00 Min. :0.0000   
## 1st Qu.: 23.00 1st Qu.:0.0000   
## Median : 47.00 Median :0.0000   
## Mean : 65.33 Mean :0.2493   
## 3rd Qu.: 88.00 3rd Qu.:0.0000   
## Max. :336.00 Max. :1.0000   
## NA's :2129

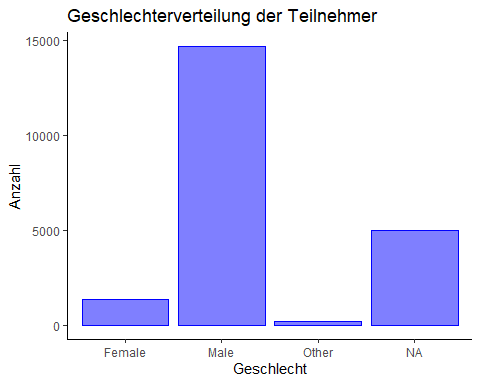
dim(datFull)

## [1] 21287 14

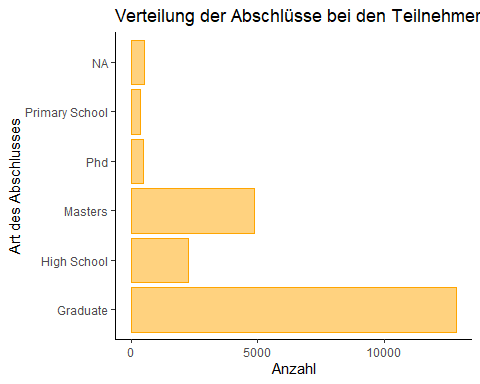
#count NAs in column gender  
sum(is.na(datFull$gender))

## [1] 5016

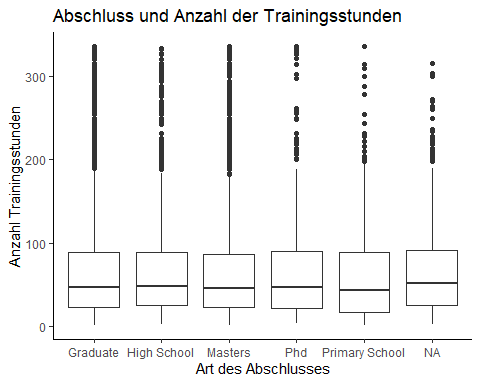
#Plot zu Geschlechterverteilung  
ggplot(data=datFull, aes(x=gender))+  
 geom\_bar(color='blue', fill='blue', alpha=.5)+  
 theme\_classic()+  
 labs(title="Geschlechterverteilung der Teilnehmer")+  
 xlab("Geschlecht")+  
 ylab("Anzahl")



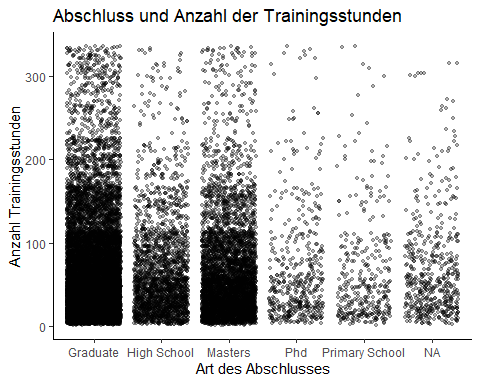
#Plot der Abschlüsse  
ggplot(data=datFull, aes(x=education\_level))+  
 geom\_bar(color='orange', fill='orange', alpha=.5)+  
 coord\_flip()+  
 theme\_classic()+  
 labs(title="Verteilung der Abschlüsse bei den Teilnehmer")+  
 xlab("Art des Abschlusses")+  
 ylab("Anzahl")



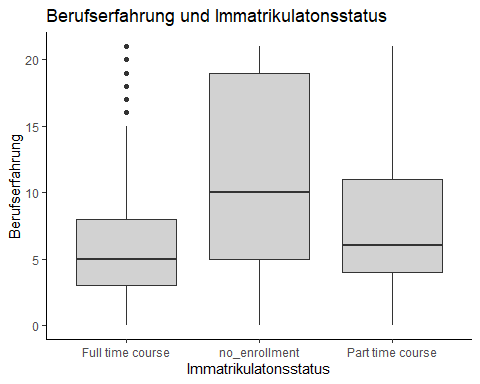
#Plot Anzahl der Trainngsstunden und Abschluss (Boxplot)  
ggplot(data=datFull, aes(x=education\_level, y=training\_hours))+  
 geom\_boxplot()+  
 theme\_classic()+  
 labs(title="Abschluss und Anzahl der Trainingsstunden")+  
 xlab("Art des Abschlusses")+  
 ylab("Anzahl Trainingsstunden")



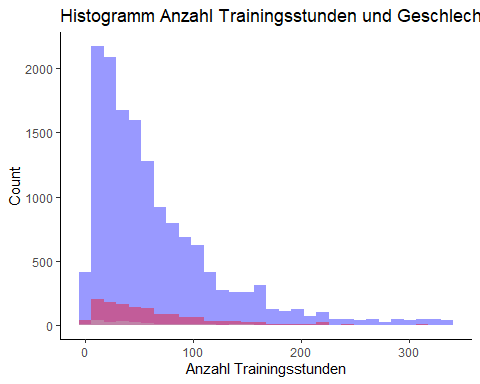
#Plot Anzahl der Trainngsstunden und Abschluss (Point/Jitter)  
ggplot(data=datFull, aes(x=education\_level, y=training\_hours))+  
 geom\_jitter(color='black', alpha=.3, size=.9)+  
 theme\_classic()+  
 labs(title="Abschluss und Anzahl der Trainingsstunden")+  
 xlab("Art des Abschlusses")+  
 ylab("Anzahl Trainingsstunden")



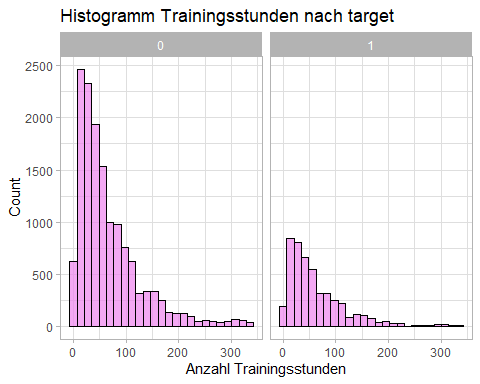
#ersetze < und > durch feste numerische Werte um typecast durchzuführen  
datFull$experience <- sub("<1", "0", datFull$experience)  
datFull$experience <- sub(">20", "21", datFull$experience)  
datFull$experience <- as.numeric(datFull$experience)  
  
datFull %>%   
 select(enrolled\_university, experience) %>%  
 filter(!is.na(enrolled\_university),!is.na(experience)) %>%  
 ggplot(data=., aes(x=enrolled\_university, y=experience))+  
 geom\_boxplot(fill="grey", alpha=.7)+  
 theme\_classic()+  
 labs(title="Berufserfahrung und Immatrikulatonsstatus")+  
 xlab("Immatrikulatonsstatus")+  
 ylab("Berufserfahrung")



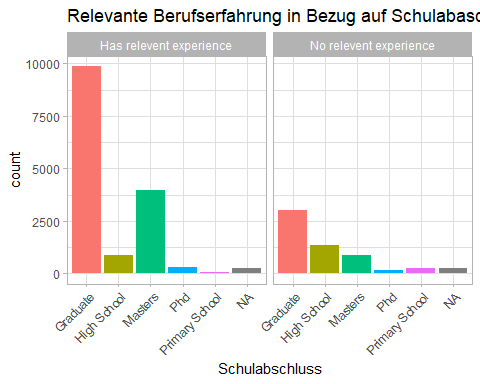
###  
ggplot()+  
 geom\_histogram(data=subset(datFull, gender == "Male"), aes(x=training\_hours), fill="blue", alpha=.4)+  
 geom\_histogram(data=subset(datFull, gender == "Female"), aes(x=training\_hours), fill="red", alpha=.4)+  
 geom\_histogram(data=subset(datFull, gender == "Other"), aes(x=training\_hours), fill="grey", alpha=.4)+  
 theme\_classic()+  
 labs(title="Histogramm Anzahl Trainingsstunden und Geschlecht")+  
 xlab("Anzahl Trainingsstunden")+  
 ylab("Count")



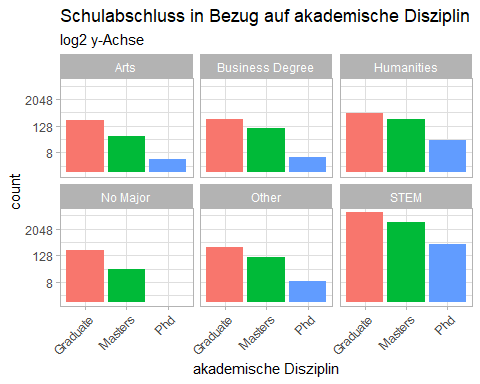
datFull %>%  
 select(training\_hours, target) %>%  
 filter(!is.na(training\_hours),!is.na(target)) %>%  
 ggplot(data=., aes(x = training\_hours)) +  
 geom\_histogram(bins = 25, color = "black", fill = "violet", alpha=.7) +  
 theme\_light() +  
 facet\_wrap(vars(target))+  
 labs(title="Histogramm Trainingsstunden nach target")+  
 xlab("Anzahl Trainingsstunden")+  
 ylab("Count")



ggplot(data=datFull, aes(x=as.factor(education\_level)))+  
 geom\_bar(data=datFull, aes(fill=as.factor(education\_level)))+  
 facet\_wrap(vars(relevent\_experience))+  
 theme\_light()+  
 theme(legend.position = "none")+  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))+  
 labs(title="Relevante Berufserfahrung in Bezug auf Schulabaschluss")+  
 xlab("Schulabschluss")



datFull %>%   
 select(major\_discipline, education\_level) %>%  
 filter(!is.na(major\_discipline),!is.na(education\_level)) %>%  
 ggplot(data=.,aes(x=education\_level))+  
 geom\_bar(aes(fill=education\_level))+  
 facet\_wrap(vars(major\_discipline))+  
 theme\_light()+  
 theme(legend.position = "none")+  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))+  
 labs(title="Schulabschluss in Bezug auf akademische Disziplin", subtitle = "log2 y-Achse")+  
 xlab("akademische Disziplin")+  
 scale\_y\_continuous(trans = 'log2')



datFull %>%   
 select(city\_development\_index, training\_hours)%>%  
 mutate(city\_development\_index = fct\_lump(as.factor(city\_development\_index), n = 20)) %>%  
 filter(!is.na(city\_development\_index)) %>%  
 ggplot(data= ., aes(x=city\_development\_index, y=training\_hours, fill = city\_development\_index))+  
 geom\_col()+  
 theme\_light()+  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))+  
 theme(legend.position = "none")+  
 labs(title="Stadtentwicklungsindex und Trainingsstunden", subtitle = "top 20 Stadtentwicklungsindexe")+  
 xlab("Stadtentwicklungsindex")+  
 ylab("Trainingsstunden")

