01 - Case Study: Pilgrim Online Banking

Nils Gandlau (Student-ID: 5467868)

15 10 2019

Contents

Data Cleaning

Renaming variables for convenience:

```r
data <- dataOriginal %>%
  rename(profit = "9Profit",
         online = "9Online",
         age = "9age",
         income = "9Inc",
         tenure = "9Tenure",
         district = "9District")

str(data)
```

```
## Classes 'data.table' and 'data.frame':   31634 obs. of  7 variables:
##  $ ID      : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ profit  : num  21 -6 -49 -4 -61 -38 -19 59 493 -158 ...
##  $ online  : num  0 0 1 0 0 0 0 0 0 0 ...
##  $ age     : num  NA 6 5 NA 2 NA 3 5 4 6 ...
##  $ income  : num  NA 3 5 NA 9 3 1 8 9 8 ...
##  $ tenure  : num  6.33 29.5 26.41 2.25 9.91 ...
##  $ district: num  1200 1200 1100 1200 1200 1300 1300 1200 1200 1100 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

Should we encode `district` as a factor? How many unique values does the feature have?

```r
unique(data$district)
```

```
## [1] 1200 1100 1300
```

Convert some columns to factor variables. This is important for regression models.

```r
data <- data %>%
  mutate(online = as.factor(online),
         age = as.factor(age),
         income = as.factor(income),
         district = as.factor(district)) %>%
  setDT()

str(data)
```

```
## Classes 'data.table' and 'data.frame':   31634 obs. of  7 variables:
##  $ ID      : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ profit  : num  21 -6 -49 -4 -61 -38 -19 59 493 -158 ...
##  $ online  : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 1 1 1 ...
##  $ age     : Factor w/ 7 levels "1","2","3","4",..: NA 6 5 NA 2 NA 3 5 4 6 ...
##  $ income  : Factor w/ 9 levels "1","2","3","4",..: NA 3 5 NA 9 3 1 8 9 8 ...
##  $ tenure  : num  6.33 29.5 26.41 2.25 9.91 ...
##  $ district: Factor w/ 3 levels "1100","1200",..: 2 2 1 2 2 3 3 2 2 1 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

Missing Values

Count the number of NAs for each variable.

```
nas <- rbindlist(lapply(names(data), function(colname) {
  nMissingValues <- sum(is.na(data[, get(colname)]))
  return(data.table(feature = colname, number_of_nas = nMissingValues))
}))

nas
```

```
##     feature number_of_nas
## 1:       ID             0
## 2:   profit             0
## 3:   online             0
## 4:      age          8289
## 5:   income          8261
## 6:   tenure             0
## 7: district             0
```

Age and income have roughly the same number of missing values: Is that coincidence, or can we say that when we don't know the age of a person, we also don't know her age? To answer this, we count the number of rows for which *both* age and income are NA.

```
data[, age_and_income_missing := ifelse(is.na(age) & is.na(income), 1, 0)]
cat("n_rows where both age and income are NA:", sum(data$age_and_income_missing))
```

```
## n_rows where both age and income are NA: 7728
```

```
data[, age_and_income_missing := NULL]
```

So yes, in most rows where either age or income are missing, both age and income are missing. Hence, we know that when we remove all rows that contain a missing value anywhere, we will remove roughly 8000 – which is not too bad considering we have 30k observations in total.

```
# Remove all rows that contain at least one NA
dataCleaned <- na.omit(data)

# Total rows removed
cat("Rows removed:", nrow(data) - nrow(dataCleaned))
```

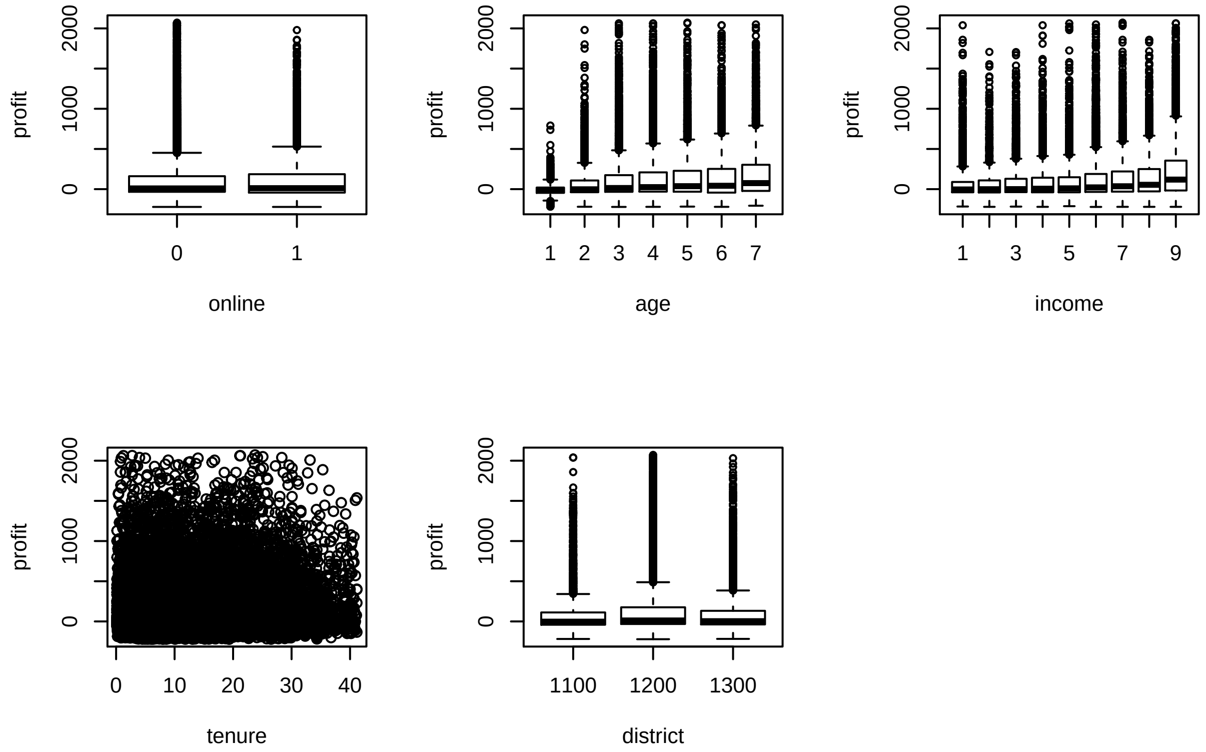
```
## Rows removed: 8822
```

Data Exploration

Plotting Features against Target

Plot each predictor against the target variable respectively. This way we get a first impression on their relationships.

```
par(mfrow = c(2, 3))
plot(profit~., data %>% select(-ID))
```

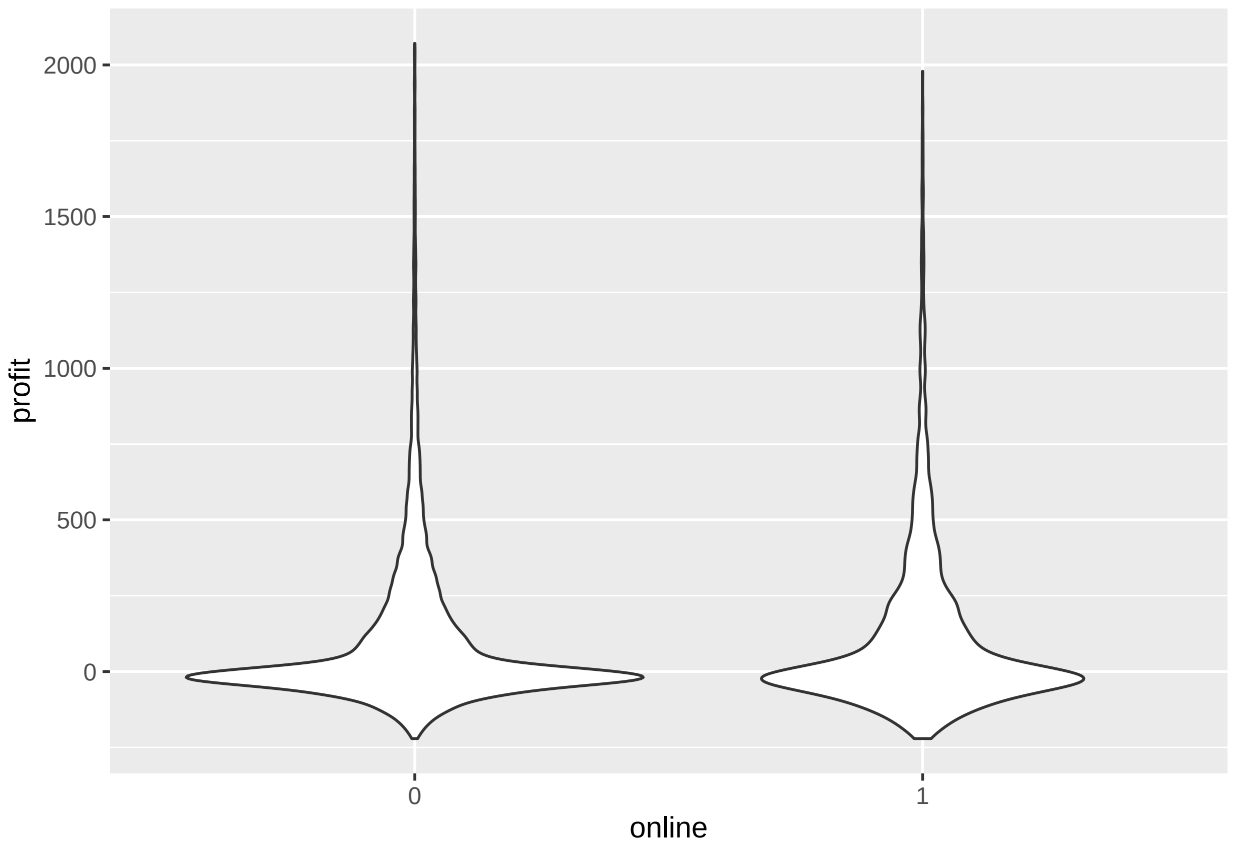


- We notice that age and income show a similar pattern, reinforcing the intution that those two predictors are probably correlated: One would think that income increases when you get older.

Distribution of profits across online/offline customers

- The profits of online & offline customers seem to resemble a pareto-distribution
- The highest-profitable customers are offline (this might be because we simply have a lot more offline customres in the data than online customers)
- The mass of slightly profitable customers seems to be larger for online customers

```
ggplot(data, aes(y = profit, x = online)) +
  geom_violin()
```

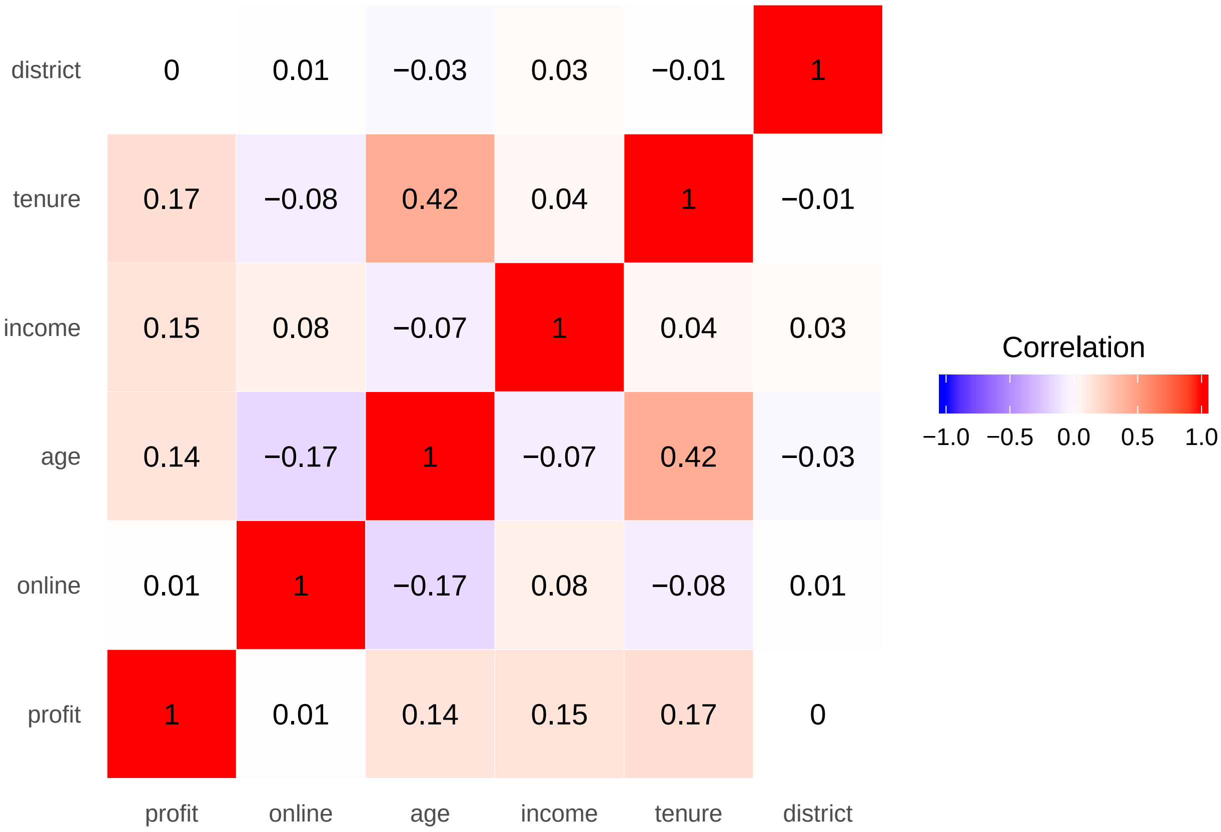


Correlation Matrix

We check for correlation across all variables, since we want to avoid the issues of **multicolinearity** (correlated predictors).

```
dataCor <- dataCleaned[, c("profit", "online", "age", "income", "tenure", "district")]
dataCor <- dataCor[, ':='(
  online = as.numeric(online),
  age = as.numeric(age),
  income = as.numeric(income),
  district = as.numeric(district)
)]

cormatrix <- round(cor(dataCor), 2)
PlotCorrelationMatrix(cormatrix)
```



	profit	online	age	income	tenure	district
district	0	0.01	−0.03	0.03	−0.01	1
tenure	0.17	−0.08	0.42	0.04	1	−0.01
income	0.15	0.08	−0.07	1	0.04	0.03
age	0.14	−0.17	1	−0.07	0.42	−0.03
online	0.01	1	−0.17	0.08	−0.08	0.01
profit	1	0.01	0.14	0.15	0.17	0

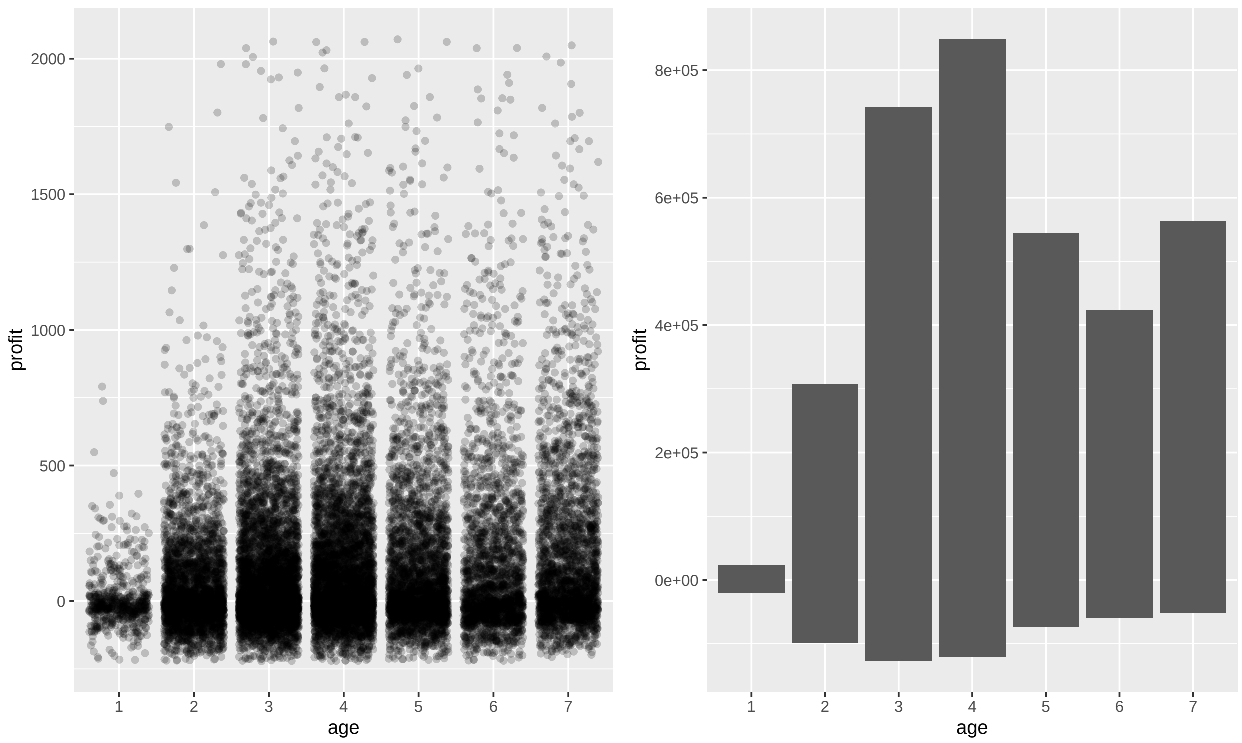
- Against intution, age seems to be only weakly correlated with income. Moreover, against intuition, it seems to be *negatively* correlated with income. Recall that correlation checks only for a linear relationship between two variables, there might be a e.g. a polynomial relationship (young people have no money, middle-aged people have a lot of money, old people again have less money).
- age is strongly positively correlated with tenure => we should avoid including both as predictors in our regression model.
- age and being online are negatively correlated (this reinforces the intuition that older people are less likely to use online banking) => We may also avoid including both as predictors
- Since `age` and `tenure` are highly correlated and `tenure` and `online` aren't, we may include `tenure` in our regression and leave age out.

Relationship between `profit` and `age`

```
p1 <- ggplot(dataCleaned, aes(x = age, y = profit)) +
  geom_jitter(alpha = 0.2)

p2 <- ggplot(dataCleaned, aes(x = age, y = profit)) +
  geom_histogram(stat = "identity")

cowplot::plot_grid(p1, p2, nrow = 1)
```



- there seems to be a non-linearity in the relationship between `profit` and `age`. If age were a continuous feature,we might have used a polynomial of degree 2 or 3 to estimate their relationship.

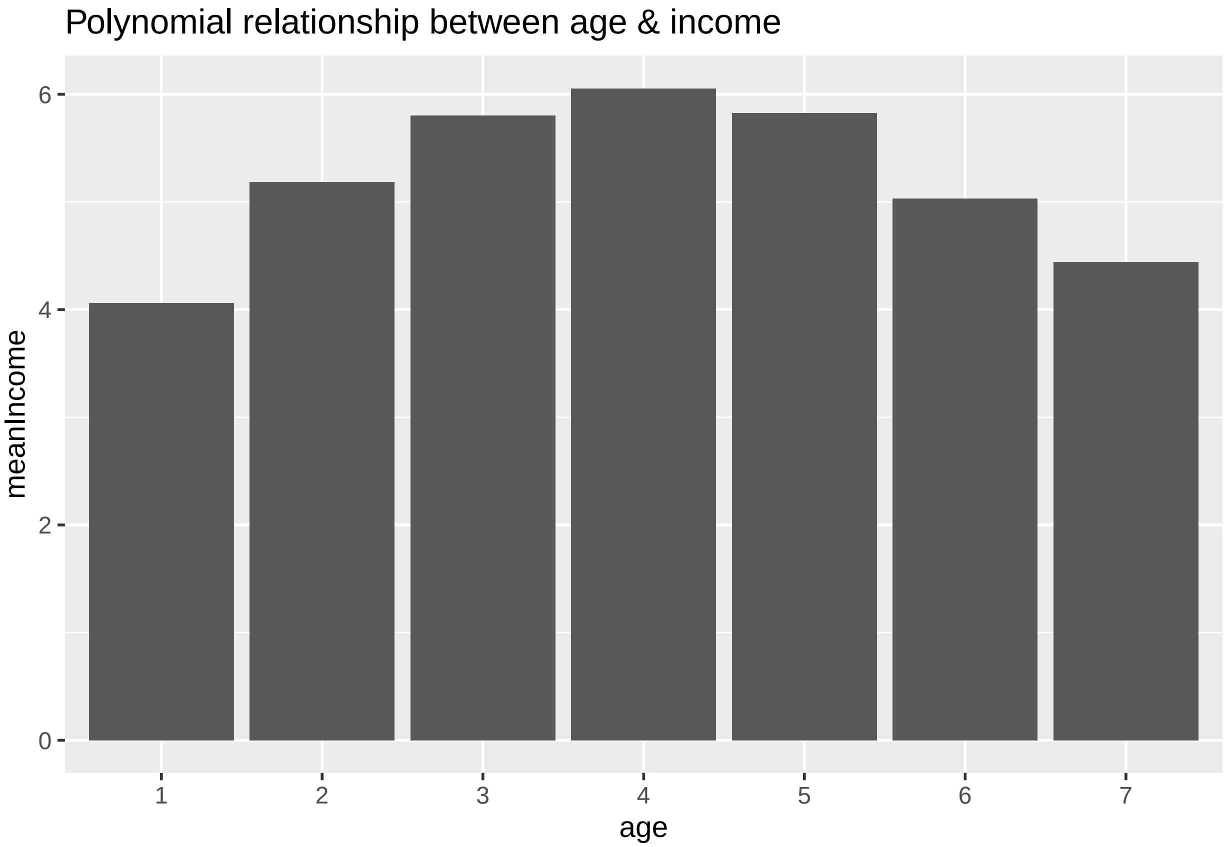
Relationship between `age` and `income`

```
dataHist <- rbindlist(lapply(unique(dataCleaned$age), function(ageCategory) {

    incomeOfAge <- dataCleaned[age == ageCategory, get("income")]
    incomeOfAge <- as.numeric(as.character(incomeOfAge))
    meanIncomeOfAge <- mean(incomeOfAge)

    return(
      data.table(
        age = ageCategory,
        meanIncome = meanIncomeOfAge
      )
    )
}))

ggplot(dataHist, aes(x = age, y = meanIncome)) +
  geom_histogram(stat = "identity") +
  ggtitle("Polynomial relationship between age & income")
```



- that's why $cor_{\text{age,income}} = -0.07$ was misleading.
- correlation-coefficient, which checks for linear relationship only, isn't able to capture the polynomial (degree-2 or possibly degree-3) relationship between age and income

Findings & Ideas

- most users of online-banking are middle-aged
- middle-aged people have the highest income
- high-income (middle-aged!) customers yield largest profits
- We have to ask whether the effect of being online changes with income – I would assume that for customers in their 20s, the effect of using online-banking is not too large, since they don't have money anyway. They might even be negative, since online-banking makes smaller, non-profitable transactions more convenient, thus driving up costs. **Thus we might consider the interaction `age * online` for the regression model.**
- Note that online-banking should in general increase interactions of customers with the bank. Thus we have to think about what type of people buy those high-margin financial products the bank offers. High-margin financial products are probably investment products, and so these are most interesting for (1) rich and (2) middle-aged people, since they have both the *financial resources* and an *available time-horizont* to benefit from interest rates.

Linear Regression

```
linreg <- lm(profit ~ online + age + income + online*age + online*income + district, data = dataCleaned)
summary(linreg)
```

```
##
## Call:
## lm(formula = profit ~ online + age + income + online * age +
##     online * income + district, data = dataCleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -499.66 -159.13  -76.16   68.20 1947.56
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -34.961     14.326  -2.440  0.01468 *
## online1           -15.677     34.682  -0.452  0.65127
## age2               29.390     13.428   2.189  0.02862 *
## age3               75.608     13.084   5.779 7.62e-09 ***
## age4               91.824     13.091   7.014 2.38e-12 ***
## age5              104.431     13.436   7.773 8.01e-15 ***
## age6              137.297     13.726  10.003  < 2e-16 ***
## age7              177.509     13.481  13.167  < 2e-16 ***
## income2             6.109     12.235   0.499  0.61760
## income3            14.748      8.861   1.664  0.09606 .
## income4            19.107      9.025   2.117  0.03426 *
## income5            25.579      9.061   2.823  0.00476 **
## income6            46.301      7.881   5.875 4.29e-09 ***
## income7            67.189      8.664   7.755 9.23e-15 ***
## income8            82.394      9.971   8.264  < 2e-16 ***
## income9           151.719      8.929  16.993  < 2e-16 ***
## district1200       19.394      6.422   3.020  0.00253 **
## district1300        7.885      7.810   1.010  0.31271
## online1:age2       23.617     30.047   0.786  0.43187
## online1:age3       33.390     29.978   1.114  0.26537
## online1:age4       33.671     30.298   1.111  0.26643
## online1:age5       58.492     32.656   1.791  0.07329 .
## online1:age6       20.153     38.555   0.523  0.60118
## online1:age7       -5.943     39.773  -0.149  0.88122
## online1:income2   -31.803     42.870  -0.742  0.45819
## online1:income3    -4.779     29.300  -0.163  0.87043
## online1:income4   -18.698     30.105  -0.621  0.53455
## online1:income5   -28.327     29.056  -0.975  0.32960
## online1:income6    -3.793     26.146  -0.145  0.88465
## online1:income7     4.893     27.414   0.178  0.85834
## online1:income8    18.775     30.022   0.625  0.53173
## online1:income9    28.531     27.222   1.048  0.29462
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 275.4 on 22780 degrees of freedom
## Multiple R-squared:  0.05327,    Adjusted R-squared:  0.05198
## F-statistic: 41.35 on 31 and 22780 DF,  p-value: < 2.2e-16
```

- The benefits of online-banking differ among different age-groups
- There is a significant positive influence on profits when middle-aged customers use online-banking:
  - middle-aged people have the both the desire and resources to invest, and therefore use financial instruments that are more profitable for the bank. Making the handling of financial instruments

easier with the help of online-banking may increase these types of interactions.
- The interaction effects of `online` and `income` indicate that online-banking might be less beneficial for customers in low-income categories. This reinforces the intuition that online-banking increases the frequency of the number of interactions between the customer and the bank, driving up costs for the bank. And the type of interactions that low-income customers make are often less profitable (e.g. small, frequent transactions)
- The bank should therefore try to get more rich and middle-aged people into using online-banking.
- The bank should avoid low-income pepole from using online-banking.

Problems

- We have seen that there is significant negative correlation between features `age` and `online` ($\rho = -0.17$). Thus, our regression coefficients might be biased due to multicolinearity.
- We note that `tenure` is highly correlated with `age` ($\rho = 0.42$). Further, `tenure` is less correlated with `online` ($\rho = -0.08$) than age is. Thus, we may repeat the same linear regression but replace `age` with `tenure`.

```r
linreg <- lm(profit ~ online + tenure + income + online*income + district, data = dataCleaned)
summary(linreg)
```

```
##
## Call:
## lm(formula = profit ~ online + tenure + income + online * income +
##     district, data = dataCleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -529.18 -153.35  -74.57   66.12 1975.67
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      11.6602     8.4793   1.375   0.1691
## online1          -6.2186    23.3341  -0.267   0.7899
## tenure            5.4421     0.2149  25.320  < 2e-16 ***
## income2           9.6229    12.2083   0.788   0.4306
## income3          11.4914     8.8344   1.301   0.1934
## income4          12.1986     8.9668   1.360   0.1737
## income5          16.2611     8.9663   1.814   0.0698 .
## income6          36.4591     7.7675   4.694 2.70e-06 ***
## income7          53.4069     8.5459   6.249 4.19e-10 ***
## income8          71.4644     9.8644   7.245 4.47e-13 ***
## income9         137.2900     8.7723  15.650  < 2e-16 ***
## district1200     14.4556     6.4028   2.258   0.0240 *
## district1300      3.5745     7.7937   0.459   0.6465
## online1:income2 -29.9795    42.7973  -0.700   0.4836
## online1:income3  -8.0274    29.2429  -0.275   0.7837
## online1:income4  -9.4272    29.9392  -0.315   0.7529
## online1:income5 -17.0106    28.8811  -0.589   0.5559
## online1:income6   7.2587    25.8951   0.280   0.7792
## online1:income7  22.9395    27.1618   0.845   0.3984
## online1:income8  31.0354    29.7171   1.044   0.2963
## online1:income9  46.7522    26.8558   1.741   0.0817 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 275 on 22791 degrees of freedom
## Multiple R-squared:  0.05555,    Adjusted R-squared:  0.05472
## F-statistic: 67.02 on 20 and 22791 DF,  p-value: < 2.2e-16
```

- using online-banking has a significant, positive effect on profits for the highest-income class. Again reinforcing the intuition that the highest-income class buys more profitable financial products, and since online-banking makes it easier for them to do so, they may buy more and thus boost profits.
- Recall that middle-aged people were the richest – hence these findings are in that sense related to our first regression.

Sanity check of findings

We can also do a rough sanity check on the hypothesis that *online-banking is most beneficial for middle-aged customers*. Below, we compute for each age category the relative amount of online (offline) customers that have yielded above-average profits (compared to their age-category's specific average).

```r
# Convert profits to a binary feature that indicates whether a customer
# yielded an above-average profit
dataCleaned[, meanProfit := mean(profit), by = age]
dataCleaned[, isAboveMeanProfit := profit >= meanProfit]

results <- rbindlist(lapply(unique(dataCleaned$age), function(ageCategory) {
  dataTemp <- dataCleaned[age == ageCategory]

  nOnline <- nrow(dataTemp[online == 1])
  nOnlineAboveMeanProfits <- nrow(dataTemp[isAboveMeanProfit == 1 & online == 1])

  nOffline <- nrow(dataTemp[online == 0])
  nOfflineAboveMeanProfits <- nrow(dataTemp[isAboveMeanProfit == 1 & online == 0])

  return(data.table(
    ageCategory = ageCategory,
    percOnlineAboveMeanProfits = round(nOnlineAboveMeanProfits / nOnline, 2),
    percOfflineAboveMeanProfits = round(nOfflineAboveMeanProfits / nOffline, 2)
  ))
}))

results[order(ageCategory)]
```

```
##    ageCategory percOnlineAboveMeanProfits percOfflineAboveMeanProfits
## 1:           1                       0.35                        0.33
## 2:           2                       0.36                        0.31
## 3:           3                       0.37                        0.32
## 4:           4                       0.39                        0.32
## 5:           5                       0.39                        0.33
## 6:           6                       0.35                        0.33
## 7:           7                       0.40                        0.35
```

- Across all age-groups, there seems to be a positive effect of using online-banking
- Note that the largest difference (in terms of percentage points) is found for middle-aged customers, with gains of 5-7 percentage points.

Further Ideas

- *Partialing out*: By regressing `income ~ age + age**2` we get information on how much age explains income. The residuals are then those variations in income that are not explained by age. We could then put these into our regression `profit ~ ...`.

- *Mediation Analysis*: A more sophisticated way to look for causal relationship between `profit` and `online`.