Predicting Household Composition by TV Viewing Behavior

Diploma of Advanced Studies in Applied Statistics at ETH Zurich

Rafael Lüchinger

11 April 2019

Contents

1 Abstract

2 Introduction

TV audience in Switzerland is measured by Mediapulse AG. A representative panel of roughly 2000 households is constantly under measurement. These homes were carefully selected by a complex sampling design and all household members have agreed to be part of the study. The TV viewing of each household member is individually recorded and detailed demographics are known for each person. This allows the market to target TV audiences by relevant characteristics like age gender and many more.

One issue with the panel approach is poor granularity. That means sometimes the system can not provide any audience figures for a specific channel or airtime. It is likely that in the Swiss population of about 3.5 million

households at least a few people are watching even exotic programs at exotic times of the day. However, out of a panel of 2000 households chances are high that no one was watching that content. This is not a bias of the measurement but poor resolution.

A solution to this problem could be the inclusion of third party data. Set-Top-Boxes (STB) of TV-provider (Swisscom, UPC, etc.) are automatically recording the TV consumption in millions of Swiss homes and the data is returned to the providers servers (return path data, RPD). There are still many issues with these data that are currently addressed.

One major issue of RPD is that the viewing data is on household level, not on individual level. Household-level data is of little use to the market. Because it gives no insight in target groups based on age and gender and alike.

It is unlikely that RPD provider will ever measure the individual viewing or survey individual demographics within the subscribers homes. Apart from region code, the only information about the home is the viewing data itself. So the question arises if it is possible to predict the household composition based on viewing behavior.

The aim of this study is to explore the possibility to predict the household composition within a household using TV viewing data. It seems to be a two-step-problem, first to find the number of household members and then to assign age and gender to the individuals.

We will use the *Mediapulse TV-Panel* and its viewing data to study the subject. For all households in the panel its composition including household size and age and sex of each person is known. For each panel home the viewing data will be aggregated to household level. Different supervised machine learning algorithms will be fed with features extracted from that household viewing data.

3 Target and Feauture

3.1 Target: Household Composition

In this study, the aim is to predict the household composition in form of the household size, e.g. the number of people living in the household. With 4388 individuals and 2006 households on our sample day the average household size is 2.19.

The variable *hhsize* is given in the demographics file of the TV raw-data. *hhsize* is not necessarily equal to the sum of individuals for the following reasons:

- children 0-2 years old are not part of the panel
- guests are part of the data but not counted for household size
- household size is counted 1, 2, ..., 5+, with 5+ meaning households with 5 or more members

Another detail is that household size is not necessarily constant over time. The number of people living in a household can change by natural reasons like birth, death, moving in or out. This fact is neglected here, we assume household size is constant over the 8 weeks period we are looking at.

Table 1: Household composition is the sociodemografic profile of a household, for example, household size, and age and gender of the householdmembers. The last 3 of our sample of 2006 households are shown. For this study the target to predict is *hhsize*.

	hh	hhsize	age_1	age_2	age_3	age_4	age_5	sex_1	sex_2	sex_3	sex_4	sex_5
2004	6200	2	63	72	0	0	0	2	1	0	0	0
2005	6201	2	71	73	0	0	0	1	2	0	0	0
2006	6204	4	52	47	17	13	0	1	2	1	2	0

3.2 Generating Features of Viewing Behavior

3.2.1 TV Viewing Data

TV viewing data comes in the form of daily text files. There are three types of files:

1. `dem`: all individuals with their demographics and daily weights
2. `view`: the TV viewing (live and time-shifted viewing)
3. `prog`: the program timetable with genre information

A commercial software allows Mediapulse and its clients to analyse this data via an easy to use software tool. I have written an R-package that allows to read and analyse the very same input raw-data and output the very same results (e.g. daily estimates the so called "Facts" like *Reach*, *Rating* or *Share*, etc.). Because the results between Software and R-package match, I am not only sure that the data processing in R is correct but also understand exactly the calculations being used.

Table 2: TV viewing raw-data (simplified). Reading example: On day `2018-01-01`, in household `2381` individuum `1` is watching channel `SRF 1` from `18:04:21` to `18:13:02`. Later that day this person switches to channel `ARTE` and is joined by another householdmember individuum `2`.

day	hh	ind	chn	start	end
2018-01-01	2381	1	SRF 1	18:04:21	18:13:02
2018-01-01	2381	1	ARTE	18:45:20	20:05:45
2018-01-01	2381	2	ARTE	18:45:20	19:45:03

Demographic information is simply joined on keys `day`, `hh` and `ind`. Program schedule is joined via an overlap join on keys `day`, `channel` and `start/end`. If a viewing statement overlaps with multiple programs, the statement gets duplicated and the `start/end` intervals needs to be cropped to the viewing interval boundaries.

3.2.2 Selecting 8 Weeks of Viewing Data

For this study, a sample day was fixed, and the viewing data of all panel member present at that day is collected 4 weeks prior and 4 weeks after that date. The sample day is the Sunday `2017-11-12` and comprises 2006 households and 4388 individuals respectively.

The period of eight weeks should be long enough to reflect individual viewing behavior. Autumn was chosen because during colder months people are watching more TV than in Summer. Also this period is free of holidays or unusual TV events (FIFA Worldcup, etc.). Within the eight weeks all seven weekdays are equally frequent. This is important as TV viewing differs significantly between weekends and workdays (Figure 2).

3.2.3 TV Viewing on Household Level

The TV raw-data described earlier shows that *Mediapulse TV data* is recorded for each individual. With RPD data however this is not the case. RPD data only provide viewing data on household level. Which person, or how many person are sitting in front of the TV set is unknown. Also, there is no demographic information accompanying RPD data. Here we study if it is possible to predict at least the number of household members if only TV viewing on household level is known, like with RPD data. To this end the *Mediapulse TV data* have to be aggregated form individual level to household level. This means, if more than one person is watching the same content, on household level, this is reflected by a single viewing statement. This household aggregation algorithm is somewhat more complex, but not further discussed here.

3.2.4 Feature Generation

Features are a set of variables that are used as input data for Machine Learning Classifiers or as predictors for statistical models. In both cases we aim to predict the target variable *hhsize*. To create features of viewing behavior the TV data on household level is summed up for each household and day, by different characteristics. Then, for each household the average across the 8 weeks was calculated. TV viewing is expressed as the duration of viewing in seconds.

The characteristics underlying the feature generation is guided by industry knowledge and intuition about TV viewing behavior we believe would carry information about the household composition, i.e:

1. Dimension time
    - weekend vs. working days
    - time of the day
2. Dimension content
    - type of channel
    - type of program genre

Figure 2, and Figure 3 in the Appendix are illustrating the effect of the dimension *time* on total TV viewing.

There are over 300 TV channels received in Switzerland. In comparison to other countries this so called *inspill* is very large. Because of the small size of Switzerland and its different linguistic regions there are many foreign channels being watched from the neighboring countries. For simplification the channels have been mapped to channel groups. There are 3 type of groups: channel type, language, and country of origin.

The program genre is a pre-specified variable in the program schedule files. Each program is categorized to one of the 14 different genres. The TV viewing data is overlapped and split up by the program schedule. The viewing duration is than summed up by each program genre within each household.

Table 3: The sets of features of TV viewing beahvior to predict household composition.

weekpart by time of day	channels	programs
day_weekend_02to06	chn_arts	prg_commercial
day_weekend_06to08	chn_generalistprivate	prg_info
day_weekend_08to11	chn_generalistpublic	prg_kids
day_weekend_11to13	chn_kids	prg_missing
day_weekend_13to17	chn_livestileindoor	prg_movie
day_weekend_17to20	chn_livestileoutdoor	prg_music
day_weekend_20to22	chn_local	prg_news
day_weekend_22to24	chn_movieseries	prg_other
day_weekend_24to02	chn_music	prg_series
day_workday_02to06	chn_nature	prg_service
day_workday_06to08	chn_news	prg_show
day_workday_08to11	chn_paytv	prg_sport
day_workday_11to13	chn_religion	prg_talk
day_workday_13to17	chn_sport	prg_trailer
day_workday_17to20	chn_foreign	
day_workday_20to22	chn_swiss	
day_workday_22to24	chn_english	
day_workday_24to02	chn_french	
	chn_german	
	chn_italian	
	chn_other	

Table 4: Final input data set. Shown are the first 6 rows and the first 5 columns. Earch of the 2006 households is an observation on rows and identified by the household ID *hh*. The householdsize, the taret to predict, is given in column *hhsize*. All 53 above mentioned features are given in the following columns. The values are the average daily viewing duration in seconds by feature on household level.

hh	hhsize	day_weekend_02to06	day_weekend_06to08	day_weekend_08to11
6	2	0.0000	0.0000	372.8750
9	4	88.3125	20.7500	621.5000
14	2	328.1250	39.2500	12.0000
20	2	1019.6667	555.1333	824.4667
21	2	607.3750	917.2500	3143.5000
35	1	72.0000	357.7500	2303.3125

4 Data Exploration

4.1 Classification versus Regression

Theoretically household size can be interpret both as five separate categories or as a scale (ordinal or proportional) ranging from 1 to 5. For this study we interpret household size as categorical variable and therefor apply classification not regression. One reason is that behind a specific household size very different types of household compositions may exist. For example a two person household could consist of an elderly couple, a two young students of the same gender, a single parent with a child, etc. Accordingly the TV viewing behavior in households of the same size may differ significantly given the possible variety of demographics profiles.

4.2 Log Transformation

Screening through the values of the 53 feature variables reveals that often the viewing duration is strongly right skewed and zero inflated (see Figure 4 in the Appendix). That means most households have a relatively low value of TV viewing duration but for a few households the viewing is relatively high. A log transformation makes the data more symmetric. Although in general for Machine Learning algorithms such a transformation is unnecessary, it should neither be harmful, and we continue with log transformed data. Because of many zero values we use the transformation 'log(x + 1)'.

4.3 Visualisation of Features

Before applying statistcal methods, we can simply visually explore if there is a predictor variable that separates well the households by housholdsize. Figure 5 in the Appendix shows some examples. For most features there is no such discrimination power apparent. But the amount of viewing kinds channels seems to separate small and bigger households.

4.4 Dimensionality Reduction

The 53 features are not a lot. Still, some of these variables may not carry much information or they are redundant to other variables (e.g. highly corellated). For an example of Dimensionality Reduction Principal Component Analysis (PCA) was chosen. PCA was calculated using the 'prcomp' package with centering but no scaling of the log transformed data matrix. All the features represent the same unit of viewing duration. Figure 8 in the Appendix shows the results of the PCA.

Table 5: First 6 PCs. The Variance in the feature marix is not easily seperated in orthogonal components. To reach 80% cumulativey explained variance the first 16 PCs would be needed.

	Standard deviation	Proportion of Variance	Cumulative Proportion
PC1	3.982955	0.29932	0.29932
PC2	2.272029	0.09740	0.39672
PC3	1.759215	0.05839	0.45511
PC4	1.517791	0.04347	0.49858
PC5	1.480727	0.04137	0.53995
PC6	1.324283	0.03309	0.57303

4.5 Probability to Correctly Classify by Chance

The probability to correctly classify by chance is 20%, if all five categories of household size were uniformly distributed (naive estimation).

```
x <- replicate(100, sample(1:5, nrow(hh.composition), replace = TRUE))
round(mean(apply(x, 2, function(y) mean(y == hh.composition$hhsize))), 2)
```

```
## [1] 0.2
```

However, we know the probability of each household size category in our sample:

```
p <- prop.table(table(hh.composition$hhsize))
c(hhsize = round(p*100, 2))
```

```
## hhsize.1 hhsize.2 hhsize.3 hhsize.4 hhsize.5
##    34.55    32.70    13.71    13.36     5.68
```

Therefore we can calculate the probability of correct classification by chance more precisely:

```
round(sum(p^2), 2)
```

```
## [1] 0.27
```

The estimation by simulation comes close:

```
x <- replicate(100, sample(1:5, nrow(hh.composition), replace = TRUE, prob = p))
round(mean(apply(x, 2, function(y) mean(y == hh.composition$hhsize))), 2)
```

```
## [1] 0.27
```

4.6 Partitioning into Train and Test Datasets

The 2006 households are split into two datasets, one for training and one for testing. The training data makes 60% of households and will be used to train the different models. The test data consists of the other 40% of households and will be used to test how good the trained models perform on classifying new data.

```
library(caret)
set.seed(999)
train <- caret::createDataPartition(pred.log$hhsize, p = .6, list = FALSE)
d <- list(train = pred.log[train,-1], test = pred.log[-train,-1])
```

A stratified random sampling is used to split the households. Stratification by household size guarantees that the ration of each household size category is equal between train and test data.

Table 6: The 2006 household are split randomly into train (60%) and test (40%) data, with stratification by household size.

	train n	test n	train %	test %
hhsize1	416	277	0.35	0.35
hhsize2	394	262	0.33	0.33
hhsize3	165	110	0.14	0.14
hhsize4	161	107	0.13	0.13
hhsize5	69	45	0.06	0.06

5 Model Specification

5.1 Multinomial Linear Model

```
library(nnet)
m.mnr <- multinom(hhsize ~ ., data = d$train, trace = TRUE, maxit = 500)
```

5.2 Support Vector Machine

```
library(e1071)
m.svm.linear <- svm(hhsize ~ ., data = d$train, kernel = "linear", cost = 1)
m.svm.radial <- svm(hhsize ~ ., data = d$train, kernel = "radial", cost = 15,
                    gamma = 0.01)
```

5.3 Random Forest

```
library(randomForest)
m.rf <- randomForest(hhsize ~ ., data = d$train, imortance = TRUE,
                     strata = d$train$hhsize,
                     sampsize = rep(min(table(d$train$hhsize)), 5))
```

6 Performance

6.1 Accuracy

```
models <- list(
  multinomial  = m.mnr,
  randomforest = m.rf,
  svm.linear   = m.svm.linear,
  svm.radial   = m.svm.radial
  )

pred <- list(
  train = as.data.frame(lapply(models, predict)),
  test  = as.data.frame(lapply(models, predict, newdata = d$test))
)

calc.acc <- function(predicted, observed) mean(predicted == observed)

tabl.acc <- rbind(
```

```
  train = sapply(pred$train, calc.acc, observed = d$train$hhsize),
  test  = sapply(pred$test,  calc.acc, observed = d$test$hhsize)
)
```

Table 7: The Accuracy in train and test dataset for the different classifiers.

	train	test
multinomial	0.56	0.42
randomforest	0.43	0.44
svm.linear	0.57	0.43
svm.radial	0.95	0.44

6.2 Cohen's Kappa

```
library(psych)
calc.kappa <- function(predicted, observed, param) {
  mx <- table(observed = observed, predicted = predicted)
  cohen.kappa(mx)[[param]]
}

tabl.kappa <- rbind(
  train = sapply(pred$train, calc.kappa, observed = d$train$hhsize, 'weighted.kappa'),
  test  = sapply(pred$test,  calc.kappa, observed = d$test$hhsize, 'weighted.kappa')
)
```

Table 8: Cohren's weighted Kappa in train and test dataset for the different classifiers.

	train	test
multinomial	0.64	0.48
randomforest	0.52	0.54
svm.linear	0.66	0.47
svm.radial	0.97	0.53

6.3 Confusion Matrix

Cohen's kappa takes into account the the probability of chance of each category. It is somewhat more

Cohen's weighted kappa punishes disagreement more as further appart the categories lie. For example, incorrectly classifying a household of size 1 as 5 is worse than 2. Using weighted kappa is somwhat contradicting to our earlier statement to interpret household size not as an ordered scale.

```
agree <- list(
  train = lapply(pred$train, calc.kappa, observed = d$train$hhsize, 'agree'),
  test  = lapply(pred$test,  calc.kappa, observed = d$test$hhsize, 'agree')
)
agree <- do.call(rbind, lapply(agree$test, as.data.frame))
agree$model <- sub('\\.\\d+', '', rownames(agree))
names(agree)[3] <- 'agreement'
```

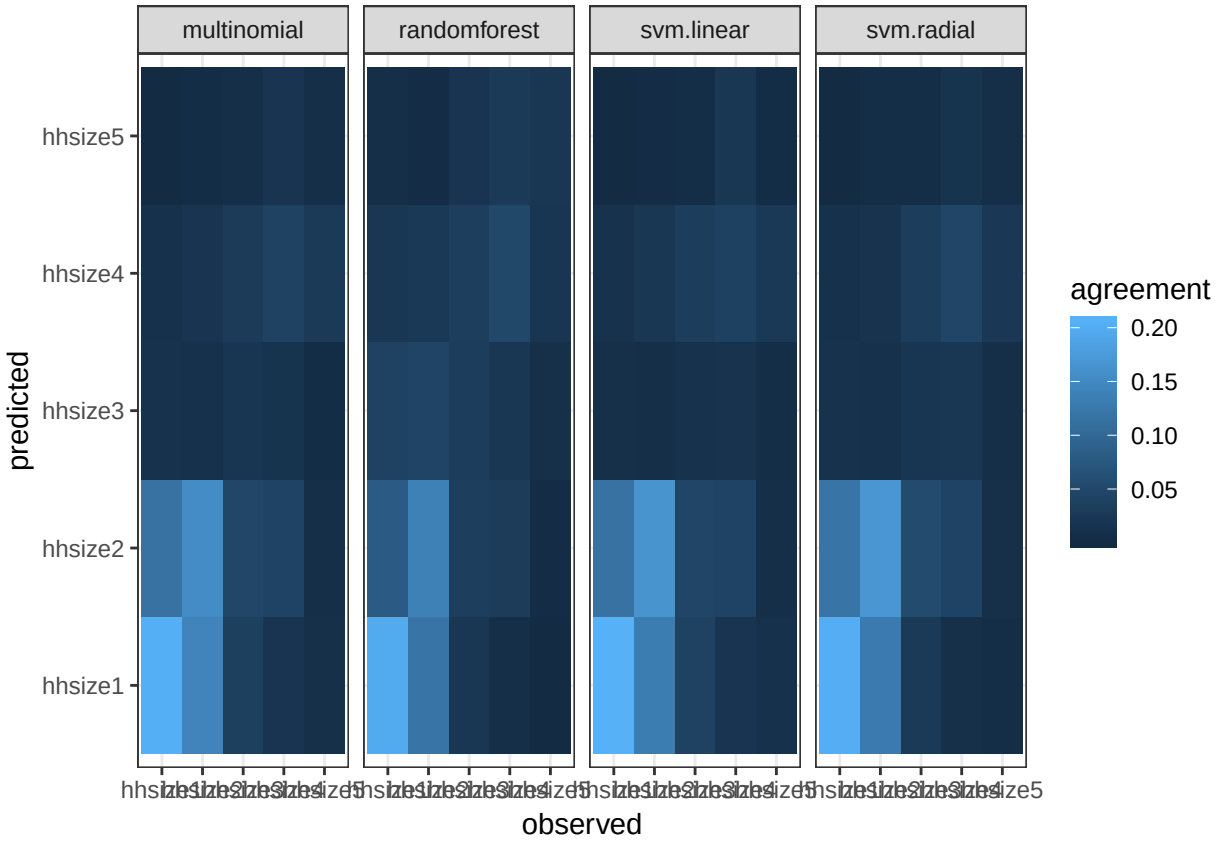


Figure 1: Heatmaps of Cohen's agreement matrix for each model

7 Discussion

8 Appendix

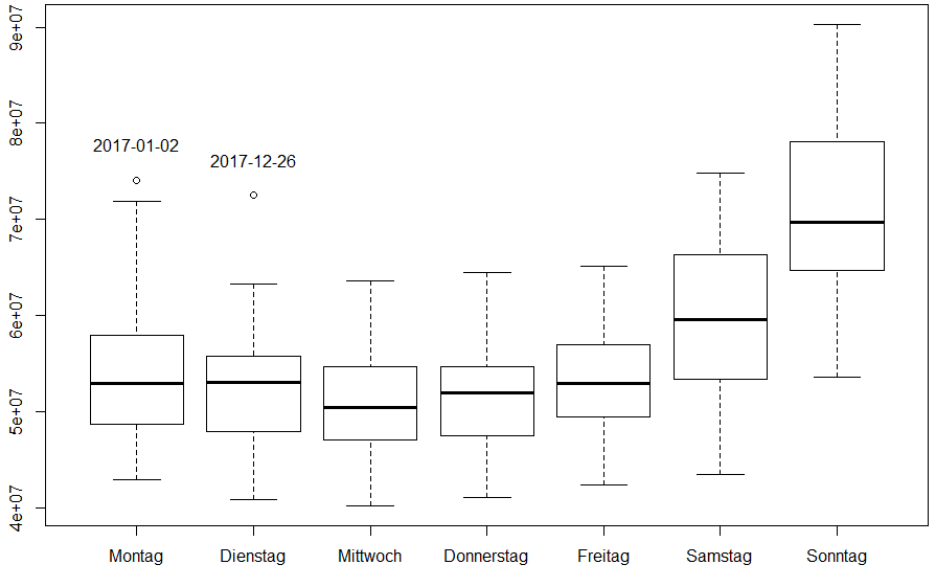


Figure 2: The sum of TV viewing duration [seconds] by weekdays during 2017. On weekends more TV is watched than during the rest of the week. Festival days often behave like Sundays.

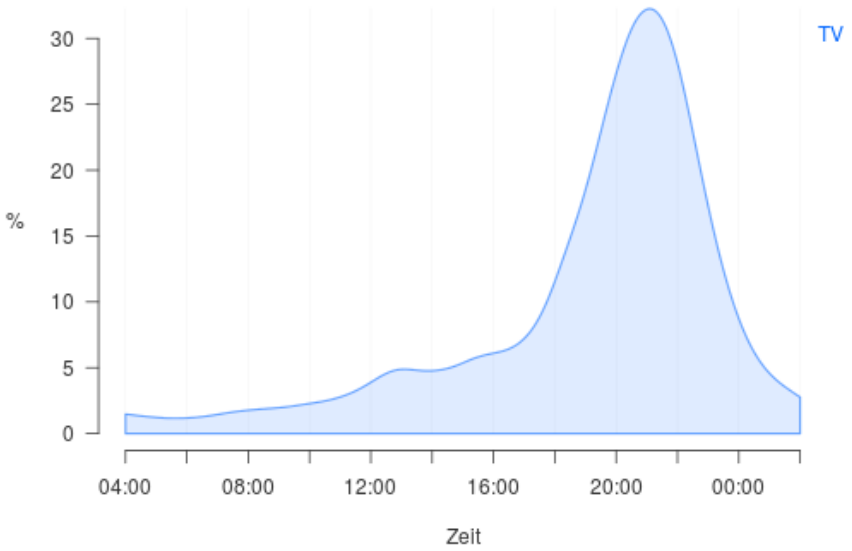


Figure 3: The relative amount of TV viewing across time of the day. The curve is the average of all 365 days in 2017. In the market the peak around 20:00 o'clock is called Primetime. On weekends the curve is flatter.

\begin{figure}[H]

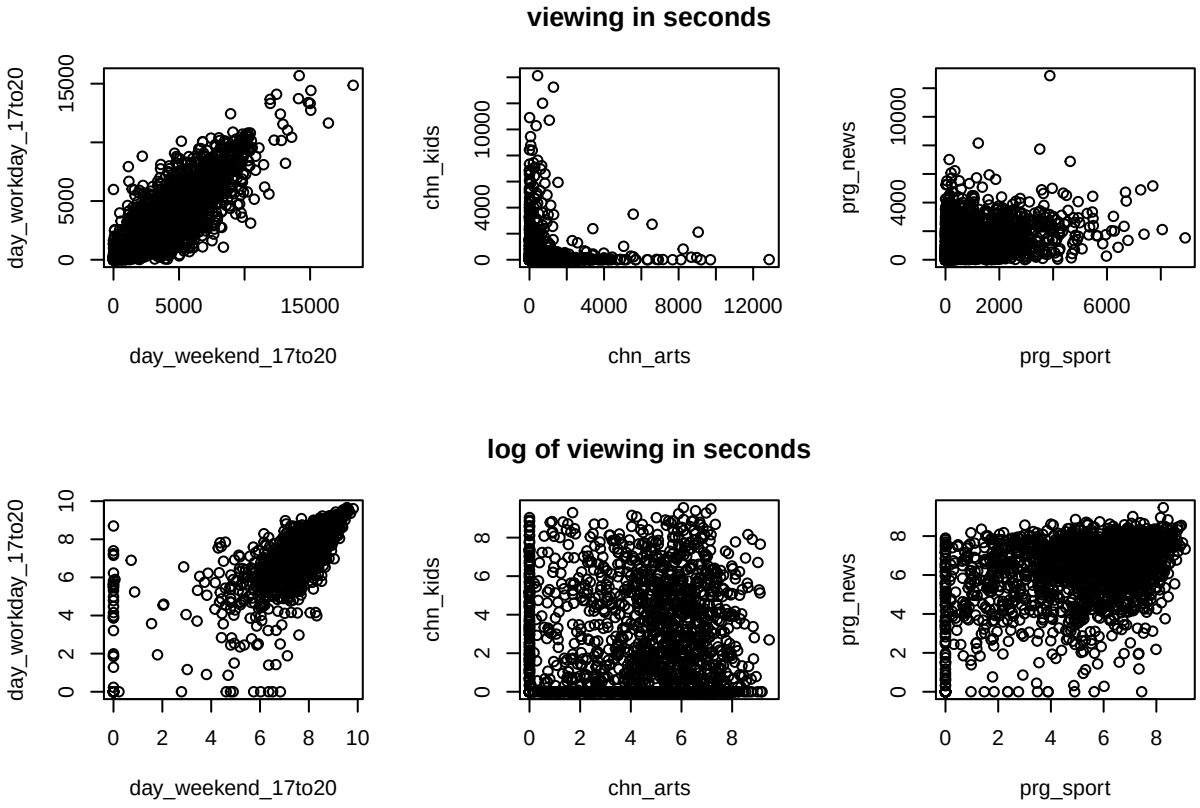


Figure 4: Illustration of log transformation. The upper row of scatterplots shows three examples of feature pairs. On of for each domain *time*, *channel* and *program*. In many cases viewing duration is not symmetrical distributed. The lower row shows the very same scatterplot with log transformed values.

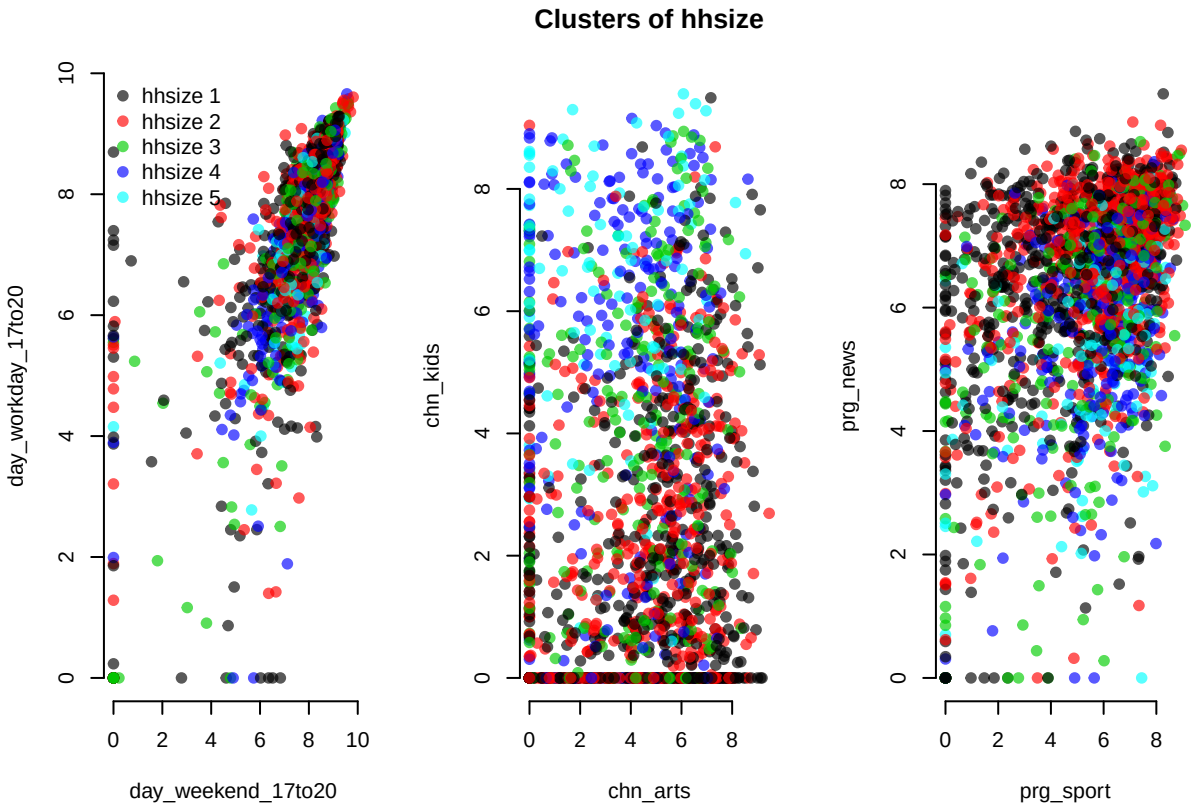
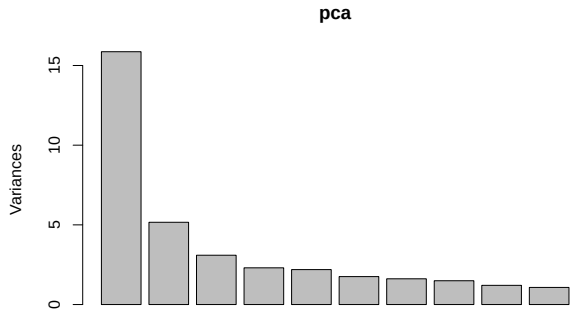


Figure 5: Shown are the same three scatterplots as in the Figure above but this time the corresponding household size is indicated by the color of the dots. If there was a feature that would separate the households (dots) into clusters of the same color, this would tell us that this particular variable is a good discriminator for household size. Apprently the variable *chn_kids* separates black and red dots from light and dark blue dots. This means there is a tendency that the more a household watches TV on typical kids channels, the more likely it is a 4 or 5 person household.



{

}

\caption{ Principal Component Analysis (PCA). Above the Screeplots shows the variance explained by the first 10 of principal components (PCs). The first PC explaine 30% of the total variance. Below, scatterplots and biplots of the first 3 PCs.} \end{figure}