Federal State Autonomous Educational Institution of High Professional Education National Research University «Higher School of Economics»

Faculty of Computer Science School of Data Analysis and Artificial Intelligence

# Report on the course "Modern Methods of Data Analysis".

«Data Sciences» Master program
Lecturer: Boris G. Mirkin.

**Authors:**
Cherny Artem
Ivanova Polina
Shvechikov Pavel

**Dataset**: Online News Popularity Data Set.
URL: http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity

## Explanation of choice

This dataset is based on recent articles, published in 2013–2015 years, so the data is actual up to now. Moreover, it is composed of different types of variables.

## Data Set Information

- The articles were published by Mashable (www.mashable.com) and their content as the rights to reproduce it belongs to them. Hence, this dataset does not share the original content but some statistics associated with it. The original content could be publicly accessed and retrieved using the provided urls.
- Acquisition date: January 8, 2015
- The estimated relative performance values were estimated by the authors using a Random Forest classifier and a rolling windows as assessment method. See their article for more details on how the relative performance values were set.
- From initial data-set we chose 34 attributes and 10000 instances (instances were chosen randomly).

**Attribute Information**:

1. `url`: URL of the article (non-predictive)
2. `timedelta`: Days between the article publication and the dataset acquisition (non-predictive)
3. `n_tokens_title`: Number of words in the title
4. `n_tokens_content`: Number of words in the content
5. `n_unique_tokens`: Rate of unique words in the content
6. `num_hrefs`: Number of links
7. `num_self_hrefs`: Number of links to other articles published by Mashable
8. `num_imgs`: Number of images
9. `num_videos`: Number of videos
10. `average_token_length`: Average length of the words in the content
11. `num_keywords`: Number of keywords in the metadata
12. `data_channel_is_lifestyle`: Is data channel 'Lifestyle'?
13. `data_channel_is_entertainment`: Is data channel 'Entertainment'?
14. `data_channel_is_bus`: Is data channel 'Business'?
15. `data_channel_is_socmed`: Is data channel 'Social Media'?
16. `data_channel_is_tech`: Is data channel 'Tech'?
17. `data_channel_is_world`: Is data channel 'World'?
18. `self_reference_avg_sharess`: Avg. shares of referenced articles in Mashable
19. `weekday_is_monday`: Was the article published on a Monday?

20. `weekday_is_tuesday`: Was the article published on a Tuesday?
21. `weekday_is_wednesday`: Was the article published on a Wednesday?
22. `weekday_is_thursday`: Was the article published on a Thursday?
23. `weekday_is_friday`: Was the article published on a Friday?
24. `weekday_is_saturday`: Was the article published on a Saturday?
25. `weekday_is_sunday`: Was the article published on a Sunday?
26. `global_sentiment_polarity`: Text sentiment polarity
27. `global_rate_positive_words`: Rate of positive words in the content
28. `global_rate_negative_words`: Rate of negative words in the content
29. `rate_positive_words`: Rate of positive words among non-neutral
30. `avg_positive_polarity`: Avg. polarity of positive words
31. `avg_negative_polarity`: Avg. polarity of negative words
32. `title_sentiment_polarity`: Title polarity
33. `shares`: Number of shares (target)

# 1 Assignment 1

In the first task we consider the attribute called `shares`, which is the number of article shares in various social networks. Let construct a histogram and boxplot of chosen attribute (see Figure 1). From the histogram wee can see, that the choosen attribute probably has a lognormal distribution, so we construct a new feautre log`(shares)`. Histogram and boxplot for this new feature log`(shares)` are presented on Figure 2.

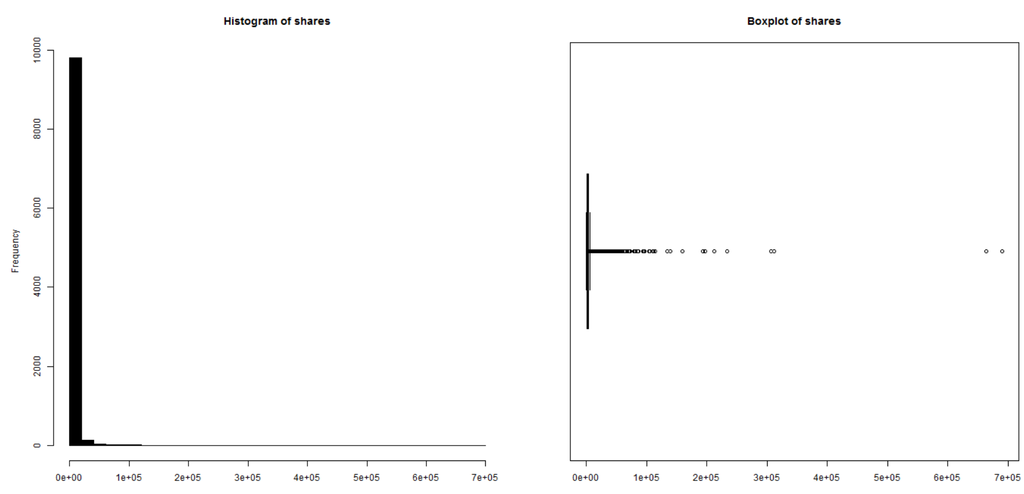



Figure 1

Below we will apply methods to the feature log`(shares)` instead of `shares`. As we can see in the Table 1, sample mean, median and mode of the log`(shares)` agree closely with each other, indicating that distribution is similar to symmetric. If we look at the same characteristics for the `shares`, we can see that the mean is significantly greater than the median and the mode. That could be easily explained with the histogram of `shares` (see Figure 1), which shows that the majority of entities have less that 20 000, and very few have more than 600 000 shares.

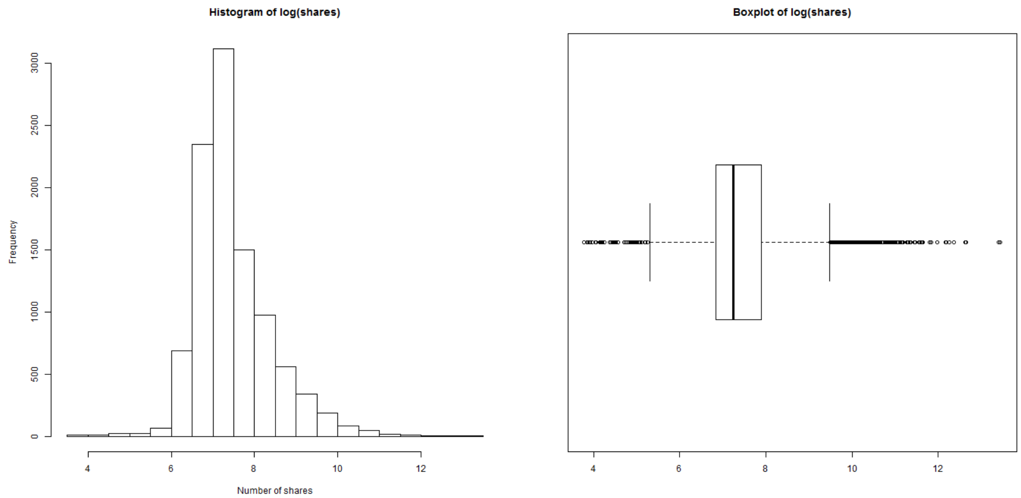



Figure 2

Table 1: Mean, median and mode for `shares` and log(`shares`)

| shares | Mean | Median | Mode |
|---|---|---|---|
| | 3374.5 | 1400 | 1100 |
| log(shares) | Mean | Median | Mode |
| | 7.45 | 7.24 | 7.0 |

## Confidence intervals for the mean

The task is to find three confidence intervals (CI) for the mean of log(`shares`). To do this, we make $N = 5000$ trials each of which consists of sampling with replacement from initial set of log(`shares`) and estimating the mean of population using that sampled data. The histogram of estimated means for the feature log(`shares`) is presented on the Figure 3, and computed 95% confidence intervals are shown in the Table 2 . The distribution of the means of log(`shares`) is very similar to normal distribution, so pivotal and non-pivotal intervals are similar too. It is worth to mention that statistic confidence interval is much more wider compared with any of the others.

## Bootstraping the mode and the median

The more the distribution resembles the power law distribution, the more appropriate is to choose median of the distribution as the center value. That is because the median is very stable against outliers. And pivotal or non-pivotal bootstrap methods can be applied to medians.

In case of mode it is hard to decide when the bootstrap technique is appropriate. The mode, in some sense, is not a smooth functional of the distribution. So the result will be most likely uninterpretable.

Table 2: 95% confidence intervals (CI's) for mean of log(`shares`)

| | |
|---|---|
| Mean | 7.45 |
| Statistic CI | (5.94; 8.97) |
| Pivotal CI | (7.35; 7.56) |
| Nonpivotal CI | (7.33; 7.58) |

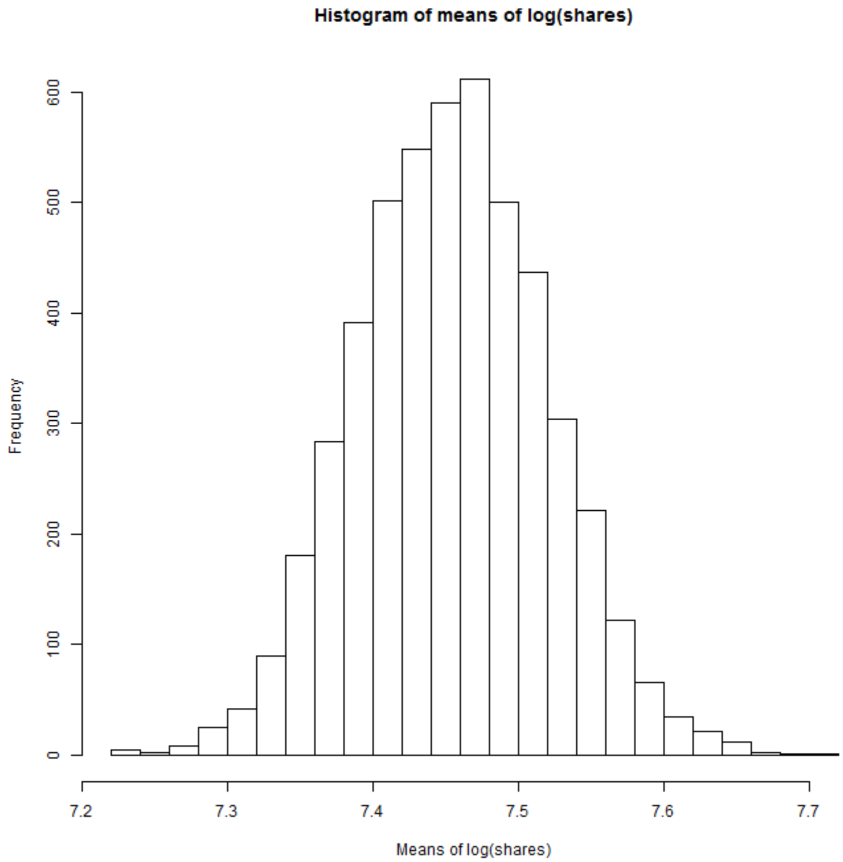



Figure 3: 30-bin histogram of means of log(`shares`).

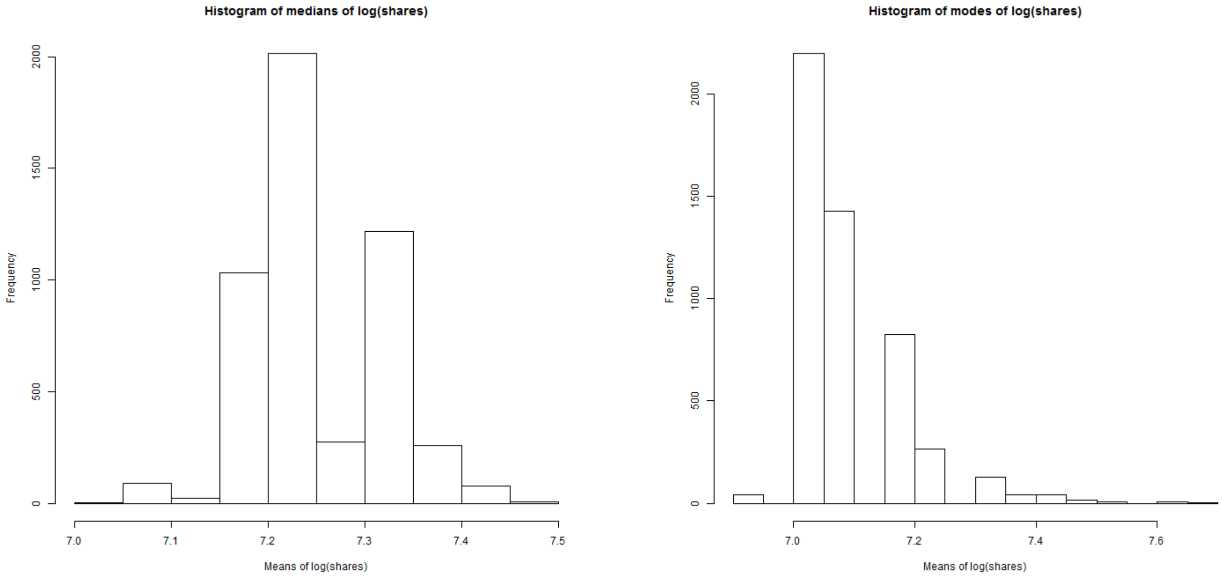
The log(`shares`) is approximately distributed normally, so the values of mean, median and mode are close to each other. Therefore in this case bootstrap is likely to be a reliable method for computing confidence intervals of median and mode.

Histograms of sample medians and modes are presented in Figure 4, and respective confidence intervals are presented in Table 3. The distribution of the modes is far from the normal random variable distribution, so pivotal confidence interval could be uncorrect. From the Table 3, as one could notice, it is evident that value of the mode is close to the left border of non-pivotal confidence interval.

It is much more interesting to compute the confidence interval for median on initial scale, that is not the median of logarithmic feature, but the median of initial `shares` feature. Since the distribution of `shares` is more similar to the power type distribution than to the normal one, it is better to choose the median as the central value due to the properties of median that were explained above. In fact, we can use either pivotal or non-pivotal approaches to estimate a median because of the next theorem.

**Theorem** (Median Theorem, [**?**])**.** *Let a sample of size $n = 2m + 1$ with $n$ large be taken from an infinite population with a density function $f(\bar{x})$ that is nonzero at the population median $\tilde{\mu}$ and continuously differentiable in a neighborhood of $\tilde{\mu}$. The sampling distribution of the median is approximately normal with mean tilde$\mu$ and variance $\frac{1}{8f(\tilde{\mu})^2 m}$.*

The histogram of sample medians of `shares` is presented at Figure 5 (a), and computed confidence intervals are shown in the Table 3. The mode's distribution is obviously far from normal (so we don't examine pivotal CI), and, as well as in the case of log(`shares`) modes, the value of population mode is close to the left border of non-pivotal confidence interval.




a) b)

Figure 4: Histograms of medians (a) and modes (b) of log(`shares`).

Table 3: 95% confidence intervals (CI's) for median and mode of log(`shares`) and of `shares`

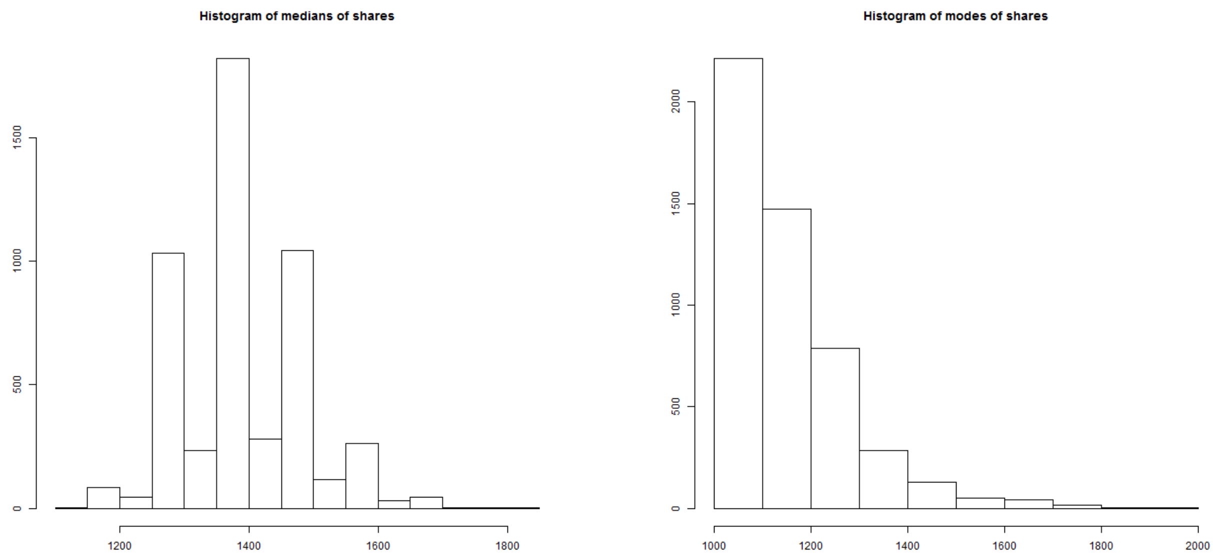
| | log(`shares`) | | `shares` | |
|---|---|---|---|---|
| | Median | Mode | Median | Mode |
| Value | 7.24 | 7.0 | 1400 | 1100 |
| Pivotal CI | (7.14; 7.36) | — | (1257.76; 1571.5) | — |
| Nonpivotal CI | (7.17; 7.38) | (7.0; 7.3) | (1250; 1600) | (1100; 1500) |

## Partitioning the population into two groups

We split our log(`shares`) data according to the day, when the article was firstly published: workday or weekend. In our dataset we have seven dummy variables, indicating the day of publishing a news: `weekday_is_monday`, `weekday_is_tuesday`, `weekday_is_wednesday`, `weekday_is_thursday`, `weekday_is_friday`, `weekday_is_saturday`, `weekday_is_sunday`. If we take entities, for which `weekday_is_saturday` or `weekday_is_sunday` are equal to 1, we will end up with the class `published_on_weekend`. All of the other entities will be considered as belonging to the class `published_on_workday`.

Histograms of the sample means in each of the classes are shown in figure 6. Each of the histrograms closely resembles the density of normal distribution, therefore pivotal and non-pivotal bootstrap methods should compute the similar confidence intervals (CI's). 95% intervals for mean in each of the two groups are presented in Table 4.

The CI of the mean in both classes do not intersect with each other, so we can claim with 95% confidence that two means in these groups are different.




a) b)

Figure 5: Histograms of sample medians (a) and sample modes (b) of `shares`.

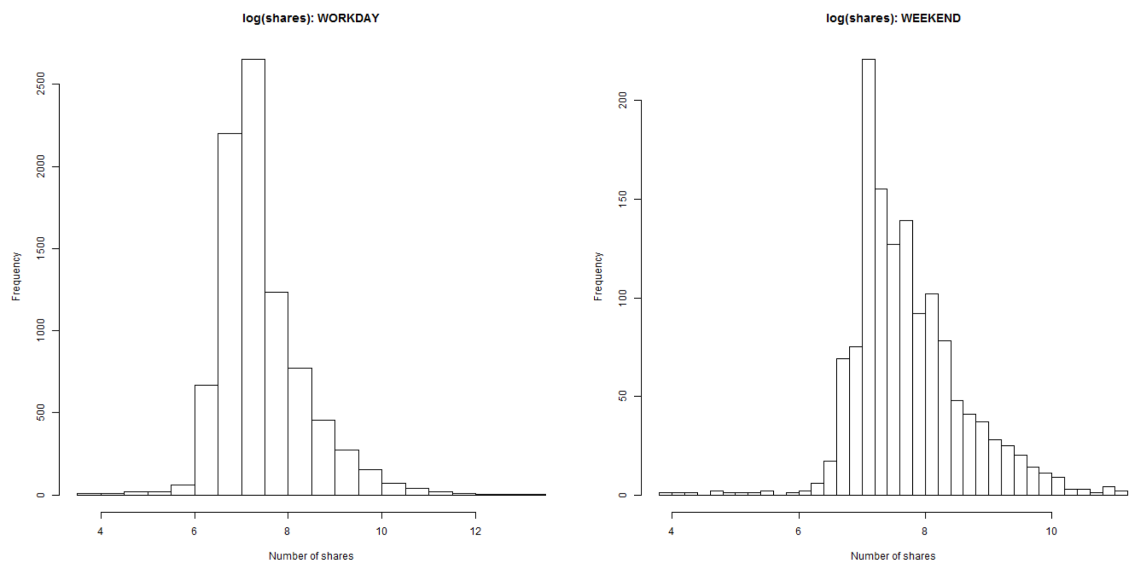



Figure 6: 30-bin histograms of log(`shares`), grouped by the day, when the article was published: workday (left) or weekend (right)

# 2 Assignment 2. Linear regression.

## 2.1 Selection of features

First of all, let's take a look at all continuous features' scatterplot to indentify which of them are linear dependent. Some features have nearly log-normal distributions, so, for more accurate and reliable linear regression we will logarithm these features.

As we can see from the scatterplot, the majority of pairs are not linear dependent. Fortunately, `global_sentiment_pola` and `rate_positive_words` are linear dependent and we can easily understand why: they measure practicaly the same characteristic. The first feature is normalized from 0 to 1 (in out data from 0 to 0.7), the second one takes

Table 4: Confidence intervals of mean of log(`shares`) feature, grouped by a day, when the article was publishead: workday or weekend

| | Workday | Weekend |
|---|---|---|
| Number of variables | 8660 | 1340 |
| Mean | 7.41 | 7.73 |
| Pivotal CI | (7.3; 7.5) | (7.63; 7.83) |
| Non-pivotal CI | (7.28; 7.54) | (7.6; 7.85) |

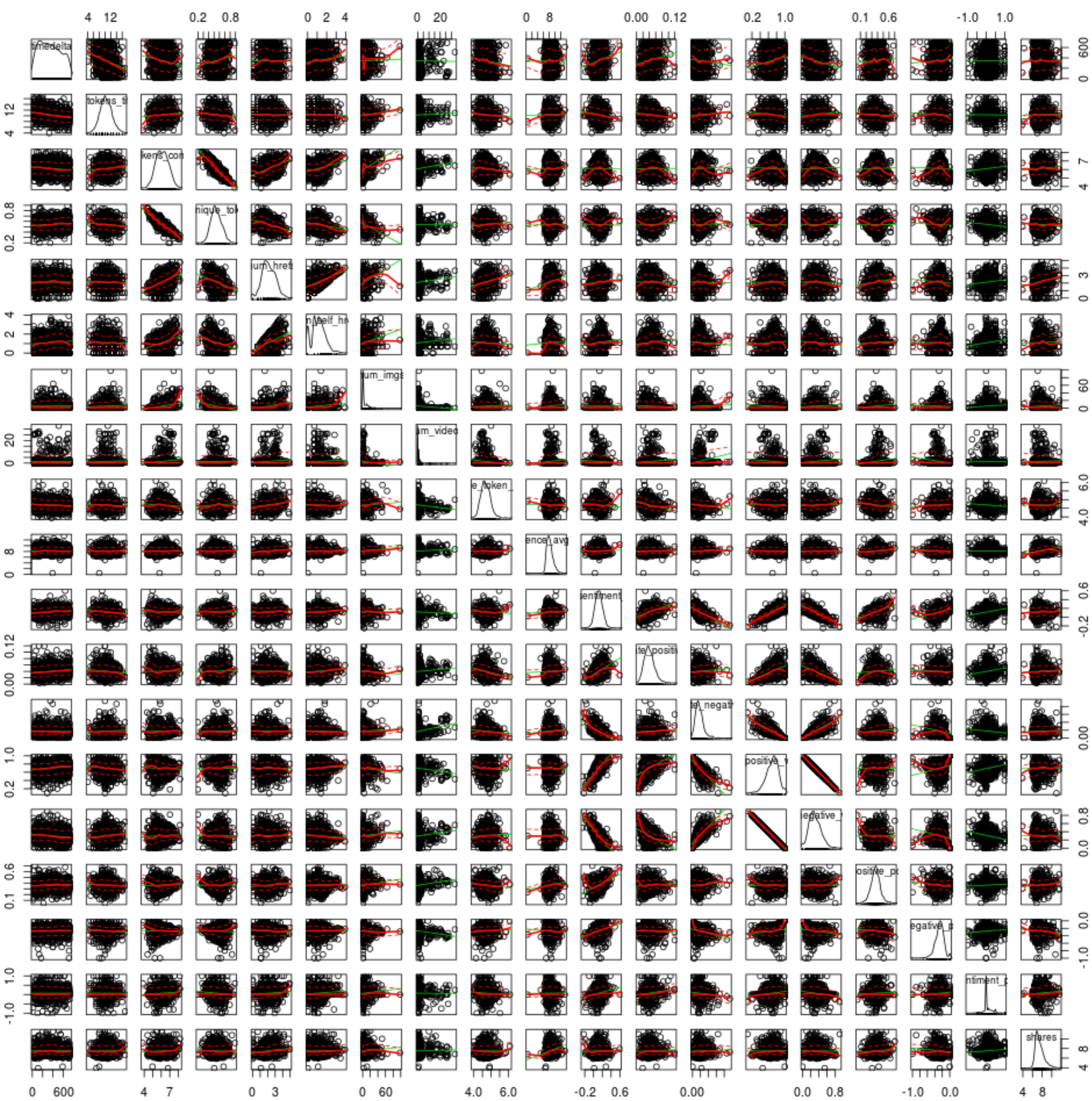
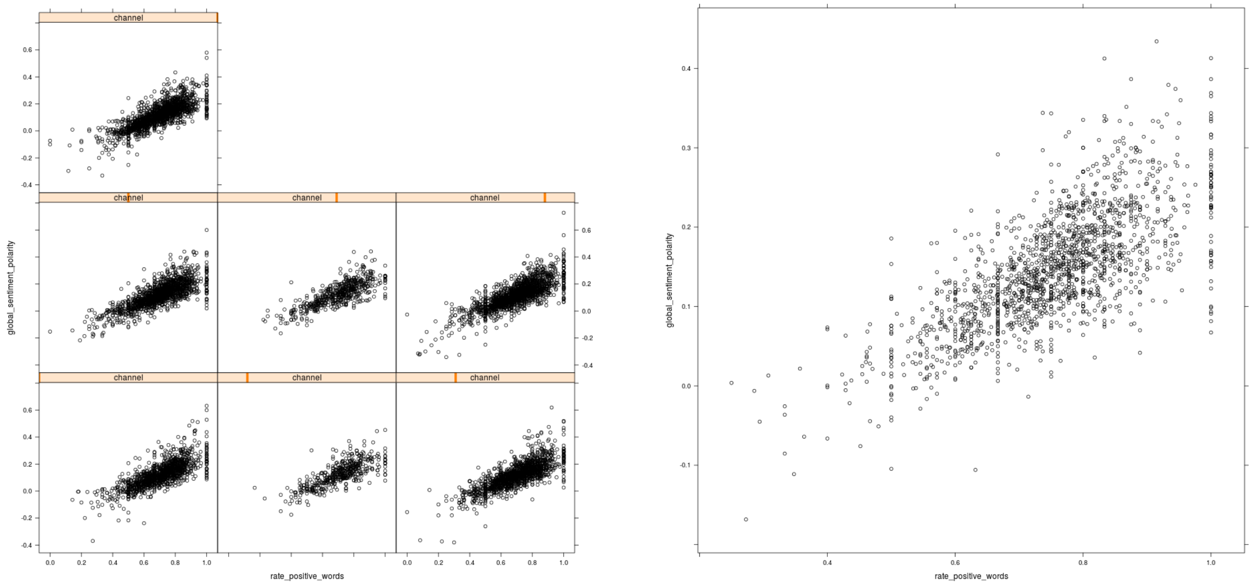


Figure 7: Scatterplot matrix of all considered continuous features

values from 0 to 1.

We will predict sentiment polarity over positive words rate. As we have rather heterogeneous data, let's make sure we won't be able to do our regression better with the help of groupping by `Channel` (Figure 8, a).

All chanells looks very similar, and we dicided to consider only technical channel (just to reduce the sample size). After all these actions our scatterplot looks like at (Figure 8, b). Further at this section we will call the




a) b)

Figure 8: Groupped by channel (a) and only technical channel (b) dependece between `global_sentiment_polarity` and `rate_positive_words`.

predicted and the prediction features just `global_sentiment_polarity` and `rate_positive_words`, implying we work with only one channel.

## 2.2 Model of linear regression

Using basic functions in R, we have built a linear regression with slope equals 0.4341 and intercept equals -0.1788. The results of the regression you can see in Figure 9.

The slope is significantly positive (p-value equals 0) and it's not surpising: the more positive words are in the article, the more text is of positive polarity.

## 2.3 Correlation and determinacy coefficients

The correlation equals 0.7170736 and the coefficient of determination equals 0.5142 (ajusted is 0.5139). As we know from the definition of the coefficient of determination, $R^2$ measures of how well the regression line approximates the real data points and equals the ratio of explained variance. It's believed in practice that $R^2 > 0.5$ is acceptable, but not enough accurate.

Particularly the value of 0.5139 means that about 51% of variablility between the two `global_sentiment_polarity` and `rate_positive_words` is captured by the linear model built with linear regression and the remaining 49% of variability still remains unaccounted for.

In another words the value of determination coefficient $R^2$ shows the rate of decrease of the variance of `global_sentiment_polarity` after its linear relation to `rate_positive_words` has been taken into account by the regression.

## 2.4 Bootstrap

We have conducted 5000 boostrap trials to estimate 95% confidence intervals of slope, intercept and correlation coefficient. The results are summarized by the histograms shown in the figures 10-12.

It can be easily seen that histograms are pretty similar to the normal type of distribution. But let us prove that.

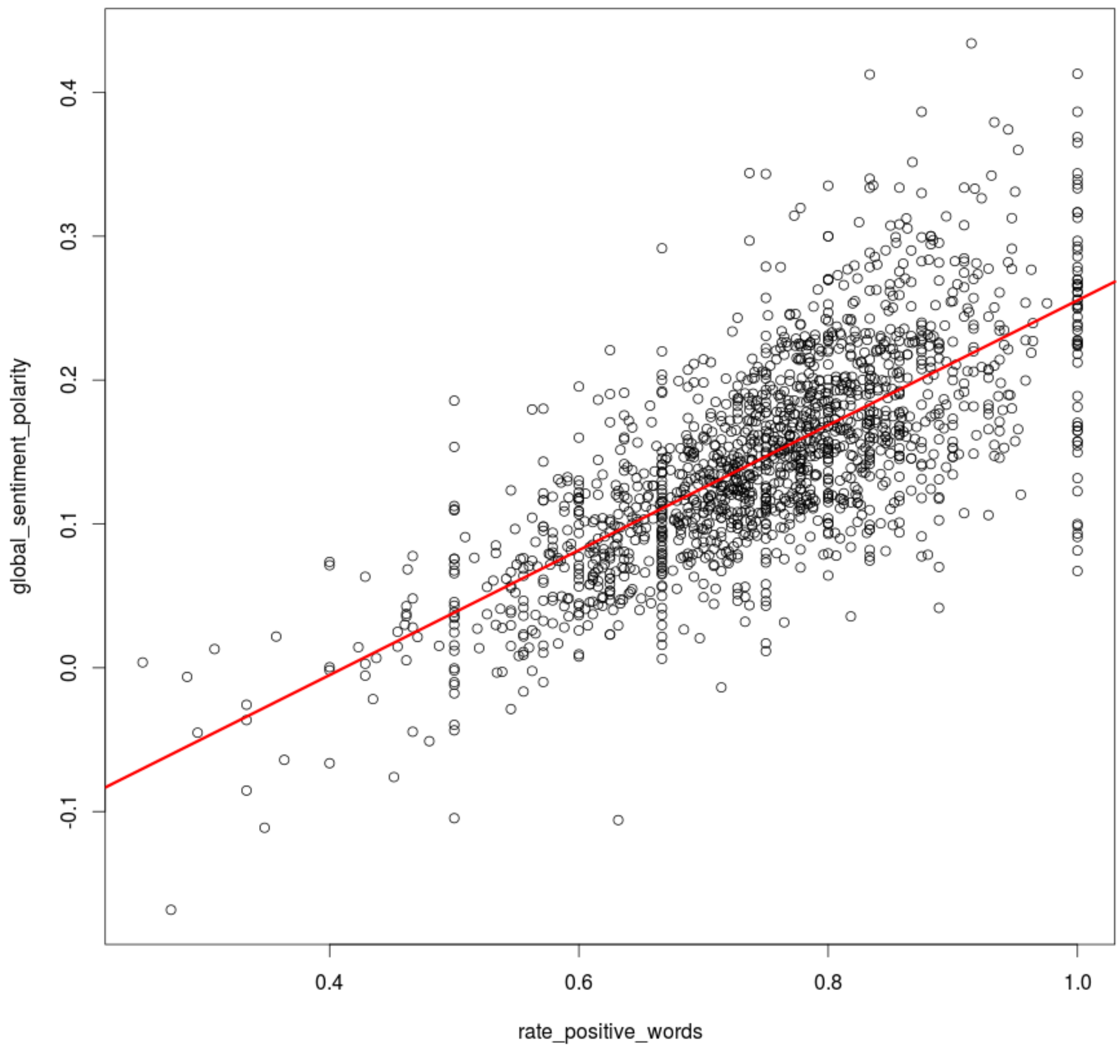



Figure 9

We have performed the Shapiro-Wilk normality test with intercept, slope and correlation coefficient bootstrap distributions. The results shown in the Table 5 are a little bit striking: with good p-value it is trustworthy that intercept and slope boostrap samples are obtained from the normal distribution, but we could not say so about correlation coefficient samples data.

So for the correlation coefficient we compute CI using ranked quntiles. The aforesaid results we obtained are shown in Table 6. It is worth to mention that the difference in estimating correlation coefficient CI using the assumption of normallity and without such is not very huge, but we think it is a good idea not only test normality of the data by it's visualization but also with statistical tests.

Table 5: Shapiro-Wilk normallity test p-value

| | 95% p-value |
|---|---|
| Intercept | 0.2279 |
| Slope | 0.244 |
| Correlation | 5.621e-08 |

Table 6: 95% confidence intervals (CI's) for intercept, slope, correlation coefficient based on bootrap techinque

| | 95% CI |
|---|---|
| Intercept | (-0.209; -0.148) |
| Slope | (0.392; 0.476) |
| Correlation (normal) | (0.668; 0.765) |
| Correlation (percentile) | (0.666, 0.763) |

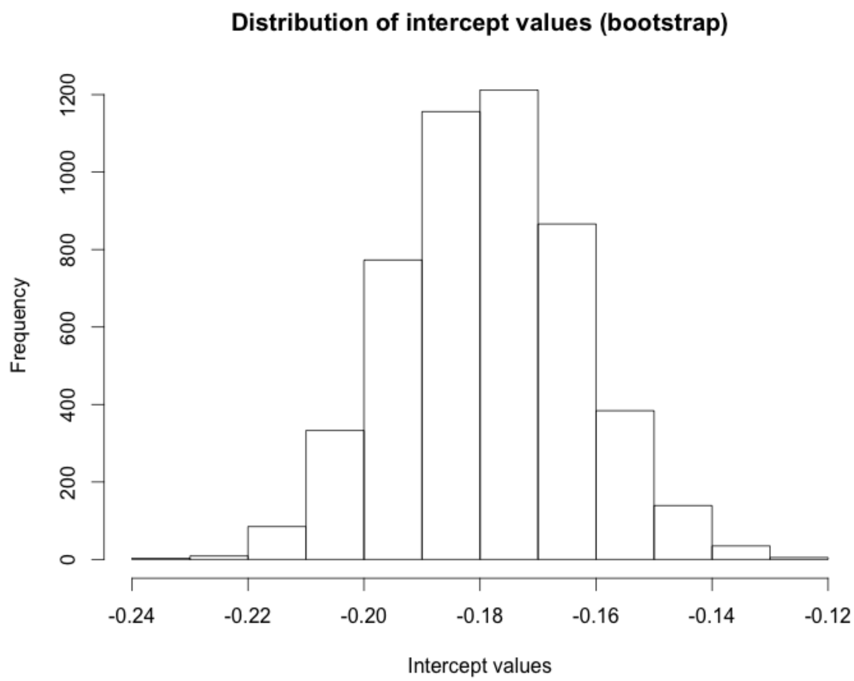



Figure 10: Distribution of intercept value of linear regression models built with each pair of sampled (`global_sentiment_polarity`) and (`rate_positive_words`) using bootrsap

## 2.5 Average relative error

Recall average relative error (ARE) and coefficient of determination ($\mathrm{R}^2$) definitions:

$$\mathrm{ARE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|,$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

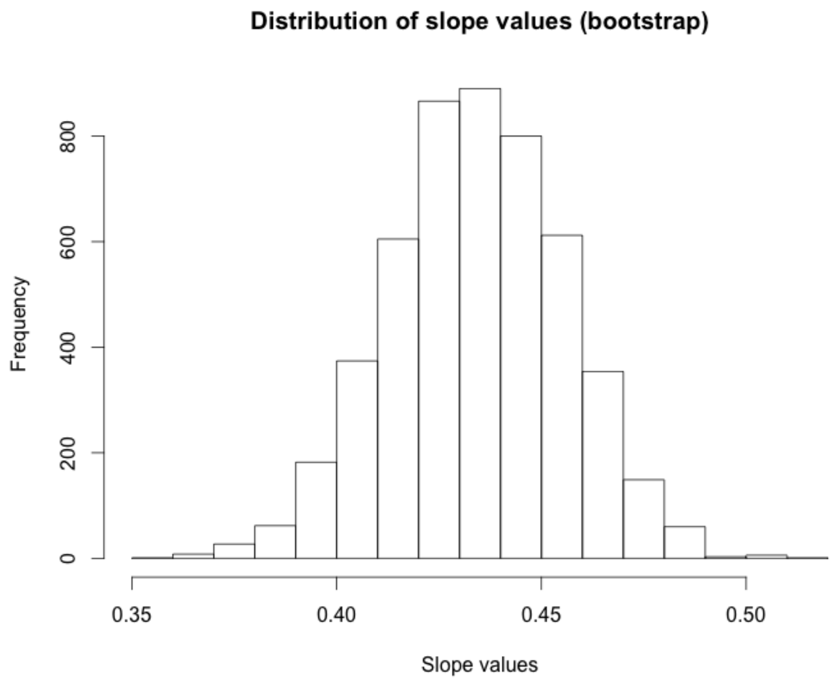



Figure 11: Distribution of slope value of linear regression models built with each pair of sampled (`global_sentiment_polarity`) and (`rate_positive_words`) using bootrsap

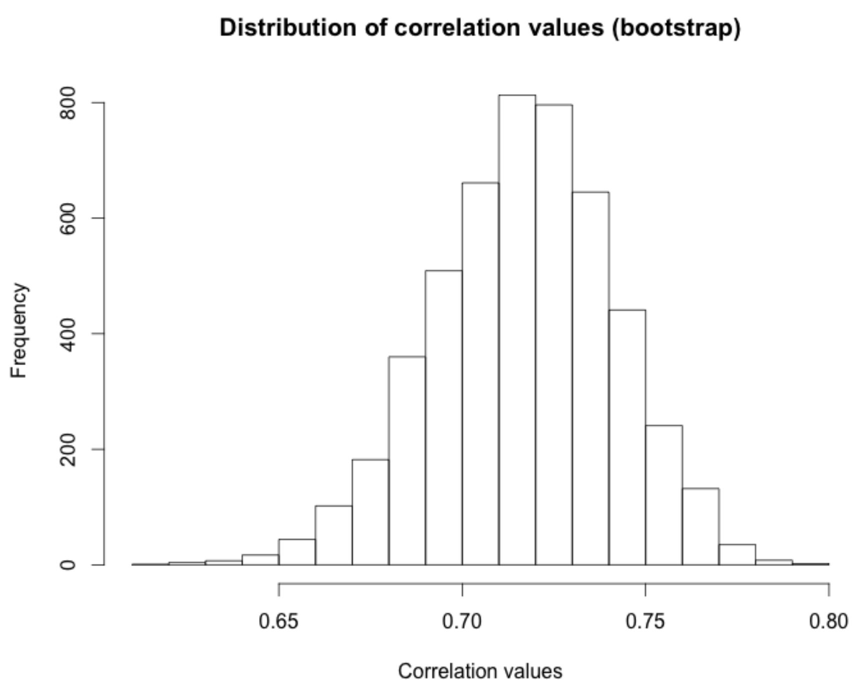



Figure 12: Distribution of correlation values between dependent variable (`global_sentiment_polarity`) and regressor (`rate_positive_words`)

```
mean( abs((y.feature - model$fitted.values)/y.feature) * 100) # in %

## [1] 56.3725

summary_regression$r.squared * 100 # in %

## [1] 51.41945
```

As we can see, considered values are reasonably close. But we should note that ARE is sensitive to the addition of a constant to all of the $y_i$ while $\mathrm{R}^2$ is not. That is, we could obtain any arbitrary value of ARE keeping the coefficient of determination constant by adding some constants to $y_i$.

It was suprise to us that ARE can be greater than 1 (which is the case if $y_i$ is much less than $y_i - \hat{y}_i$).

According to all these facts one could conclude that comparing ARE and $\mathrm{R}^2$ is meaningless without additional assumptions.

## 2.6 Nature-inspired algorithm

We will use nature-inspired algorithm to compute the parameters of linear regression that minimize the absolute relative error. If the target values of regression are $y_i$ and predicted by linear regression values are $\hat{y}_i$, the absolute relative error is:

$$\frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|. \tag{1}$$

We have implemented algorithm similar to the one, which is described in [**?**], but we will minimize the function `delta(coefficients, x, y)`, which looks like

```
delta <- function(coefficients, x,y){
  a = coefficients[1]; b = coefficients[2]
  yp <- a*x + b
  esq <- mean( abs((y - yp)/y) )
}
```

We also need the function that compute permissible limits for coefficients. Let look at two values of target $y_i = a \cdot x_i + b$ and $y_j = a \cdot x_j + b$ $(i \neq j)$ and express $a$, $b$ in terms of $x$ and $y$:

$$a_{ij} = \frac{y_j - y_i}{x_j - x_i}, \; b_{ij} = \frac{y_i x_j - y_j x_i}{x_j - x_i}.$$

And then we calculate max and min of coefficients $a$ and $b$ among all pairs $(x_i, y_i)$ and $(x_j, y_j)$.

Using the nature-inspired approach (we have implemented it in R-function nlr) we obtained the following values of slope $(a)$ and intercept $(b)$ and value of relative error:

```
# The regression model: y.feature = a * x.feature + b
# Coefficients of slope and intercept respectively:
model.nlr <- nlr(x.feature, y.feature)
model.nlr

## [1]  0.3610875 -0.1570541

# Value of relative error:
eps.nlr <- y.feature - model.nlr[1]*x.feature - model.nlr[2]
mean( abs(eps.nlr / y.feature) ) * 100

## [1] 50.04851
```

Comparing the values of two relative errors, we can see that nature-inspired approach reduce its value by a few percent.

In R-language there is a package `genalg` with function `rbga` that implement this approach. The results obtained with function `rbga` are very close to the results described above:

```
# The regression model: y.feature = a * x.feature + b
# Calculate the permissible limits for a and b
bound <- ddr(x.feature, y.feature)
bounds.min <- c(bound[[1]][1],bound[[2]][1]) # (a.min, b.min)
bounds.max <- c(bound[[1]][2],bound[[2]][2]) # (a.max, b.max)

rbga.res <- rbga(bounds.min, bounds.max, popSize = 30, iters = 5000,
   evalFunc = function(coefs) delta(coefs, x.feature, y.feature))

# Results (we need to take a look at "Best Solution")
cat(summary(rbga.res))

## GA Settings
##   Type                  = floats chromosome
##   Population size       = 30
##   Number of Generations = 5000
##   Elitism               = 6
##   Mutation Chance       = 0.333333333333333
##
## Search Domain
##   Var 1 = [-6.674999999946,6.47188552188199]
##   Var 2 = [-5.27885519411197,5.97499999995278]
##
## GA Results
##   Best Solution : 0.348751377798375 -0.14645601290741

# Value of relative error:
delta(c(0.3477, -0.1463), x.feature, y.feature) * 100

## [1] 49.92909
```

# 3 Assignment 4

## 3.1 Selection and building nominal features

In our dataset we have several binary features, such as weekdays (`weekday_is_monday`, `weekday_is_tuesday` and so on) and belonging to one of the channels (`data_channel_is_lifestyle`, `data_channel_is_entertainment` and so on). Therefore we built two nominal features:

- channel: integer values ranging between 0 and 6 ('No channel', 'Lifestyle', 'Entertainment', 'Business', 'Social Media', 'Tech', 'World')
- weekday: integer values ranging between 1 and 7

To obtain the third nominal feature we divide the feature `timedelta` into four parts : days between the article publication and the dataset acquisition.

```
timegroup <- cut(data$timedelta, breaks = 4)
```

And we break range of values of `timedelta` into intervals of approximate equal size: $(7.28, 189]$, $(189, 370]$, $(370, 550]$, $(550, 732]$. Let us note, that analysis of cross classification between timedelta and channel is

## 3.2 Contingency tables over features

Conditional cross-classification talbes between introduced nominal fetures are obtained with R-function `table` as is shown below. Results are presented in Tables 7 – 9. Simple analysis of the aforementioned tables reveals that there aren't any conceptual associations between categories. So our analisys could go deeper and discover some hidden dependencies with techinques that were explained in lectures.

```
table(data$channel, data$timegroup)
table(data$channel, data$weekday)
```

Table 7: Cross classification of the `channel` with `timegroup`

| | (7.28,189] | (189,370] | (370,550] | (550,732] |
|---|---|---|---|---|
| 0 | 412 | 327 | 361 | 381 |
| 1 | 120 | 93 | 130 | 183 |
| 2 | 573 | 473 | 341 | 348 |
| 3 | 392 | 408 | 410 | 446 |
| 4 | 81 | 154 | 163 | 192 |
| 5 | 401 | 469 | 491 | 492 |
| 6 | 901 | 564 | 373 | 321 |

Table 8: Cross classification of the `channel` with `weekday`

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 0 | 224 | 256 | 253 | 257 | 238 | 112 | 141 |
| 1 | 80 | 95 | 92 | 85 | 69 | 46 | 59 |
| 2 | 317 | 332 | 311 | 287 | 241 | 95 | 152 |
| 3 | 277 | 293 | 371 | 319 | 245 | 57 | 94 |
| 4 | 88 | 116 | 105 | 115 | 88 | 44 | 34 |
| 5 | 316 | 372 | 358 | 344 | 244 | 125 | 94 |
| 6 | 360 | 379 | 399 | 392 | 342 | 136 | 151 |

As could be clearly seen, it is pretty cumbersome to make any conclusion about data which is presented without normalization. So, for the ease of interpretaion, the same data converted to relative frequencies by relating them to the total number of entities is presented in Tables 10 – 12.

We choose `channel` as the common feature for comparison with two other features. Motivation for such choice is as follows, there is no much sense in cross-classification subgroups of time passed from the article publishing till dataset acquisition and the day of the week it was published. This point of veiw is also supported by the Table 12, wich reveals no particular irregularities of the data, except the number of publication in a particular day of week.

Quetelet relative index tables over our nominal fetures we obtain as a result of the following function:

Table 9: Cross classification of the `weekday` with `timegroup`

| | (7.28,189] | (189,370] | (370,550] | (550,732] |
|---|---|---|---|---|
| 1 | 57 | 40 | 30 | 36 |
| 2 | 48 | 41 | 37 | 45 |
| 3 | 51 | 49 | 48 | 40 |
| 4 | 41 | 45 | 49 | 48 |
| 5 | 46 | 35 | 29 | 33 |
| 6 | 21 | 13 | 16 | 17 |
| 7 | 20 | 24 | 24 | 17 |

Table 10: Conditional frequency table over `channel` and `timegroup`

| | (7.28,189] | (189,370] | (370,550] | (550,732] | Sum |
|---|---|---|---|---|---|
| 0 | 4.12 | 3.27 | 3.61 | 3.81 | 14.81 |
| 1 | 1.20 | 0.93 | 1.30 | 1.83 | 5.26 |
| 2 | 5.73 | 4.73 | 3.41 | 3.48 | 17.35 |
| 3 | 3.92 | 4.08 | 4.10 | 4.46 | 16.56 |
| 4 | 0.81 | 1.54 | 1.63 | 1.92 | 5.90 |
| 5 | 4.01 | 4.69 | 4.91 | 4.92 | 18.53 |
| 6 | 9.01 | 5.64 | 3.73 | 3.21 | 21.59 |
| Sum | 28.80 | 24.88 | 22.69 | 23.63 | 100.00 |

```r
getQueteletIndex <- function(v1, v2) {
  size <- length(v1)
  cont.table  <- table(v1, v2)
  row.sums   <- rowSums(cont.table)
  col.sums   <- colSums(cont.table)
  norm.cont.table  <-  cont.table / size
  norm.row.sums  <- row.sums / size
  norm.col.sums  <- col.sums / size
  list(QueteletIndexMatrix = norm.cont.table / (norm.row.sums %*% t(norm.col.sums)) - 1,
       PearsonIndexMatrix = (-norm.row.sums%*%t(norm.col.sums) + norm.cont.table ) /
                              sqrt(norm.row.sums%*%t(norm.col.sums)))
}
```

The results (in percent) are presented in Tables 13, 14.

As we can see from the Table 13, `timegroup` is dependent with `channel` in some values. For example, we observe rather big Quetelet relative index between Lifestyle channel and 4th time-group. [1] In addition, we can't reject a dependence between World channel and 1st time-group. It can be caused by not random sampling or by some extra-ordinary events with great response in the world.

Table 14 provides us less surprising and slightly more predictable results: all channels are almost independent with weekdays except of Lifestyle channel and Weekend pair. This observation is easy to interpret: users visit Mashable at the weekend – period when they have more free time (compared to workdays) to dedicate time to themselves.

[1] This could be explained with the hypothesis of some major classical lifestyle articles, which were written a long time before the dataset acquisition was occured.

Table 11: Conditional frequency table over `channel` and `weekday`

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Sum |
|---|---|---|---|---|---|---|---|---|
| 0 | 2.24 | 2.56 | 2.53 | 2.57 | 2.38 | 1.12 | 1.41 | 14.81 |
| 1 | 0.80 | 0.95 | 0.92 | 0.85 | 0.69 | 0.46 | 0.59 | 5.26 |
| 2 | 3.17 | 3.32 | 3.11 | 2.87 | 2.41 | 0.95 | 1.52 | 17.35 |
| 3 | 2.77 | 2.93 | 3.71 | 3.19 | 2.45 | 0.57 | 0.94 | 16.56 |
| 4 | 0.88 | 1.16 | 1.05 | 1.15 | 0.88 | 0.44 | 0.34 | 5.90 |
| 5 | 3.16 | 3.72 | 3.58 | 3.44 | 2.44 | 1.25 | 0.94 | 18.53 |
| 6 | 3.60 | 3.79 | 3.99 | 3.92 | 3.42 | 1.36 | 1.51 | 21.59 |
| Sum | 16.62 | 18.43 | 18.89 | 17.99 | 14.67 | 6.15 | 7.25 | 100.00 |

Table 12: Conditional frequency table over `weekday` with `timegroup`

| | (7.28,189] | (189,370] | (370,550] | (550,732] | Sum |
|---|---|---|---|---|---|
| 1 | 5.7 | 4.0 | 3.0 | 3.6 | 16.3 |
| 2 | 4.8 | 4.1 | 3.7 | 4.5 | 17.1 |
| 3 | 5.1 | 4.9 | 4.8 | 4.0 | 18.8 |
| 4 | 4.1 | 4.5 | 4.9 | 4.8 | 18.3 |
| 5 | 4.6 | 3.5 | 2.9 | 3.3 | 14.3 |
| 6 | 2.1 | 1.3 | 1.6 | 1.7 | 6.7 |
| 7 | 2.0 | 2.4 | 2.4 | 1.7 | 8.5 |
| Sum | 28.4 | 24.7 | 23.3 | 23.6 | 100.0 |

## 3.3 $\chi^2$–summary Quetelet index

We have calculated and visualize $\chi^2$–summary according to [?]: we put Pearson's indices outside the parenthesis and it's squared value inside the parenthesis. Such presentation provides mush more information than usual $\chi^2$-statistics.

As we can see from the Table 15, the largest contribution (almost 30%!) belongs to dependence between World channel and 1st time-group (most recent articles). We can sayt that dependence is positive because the Peason's index is positive.

In contrast, in Table 15 all values are approximately equal to each other, there is no outstanding dependences. So, our first suggestion about dependence between Lifestyle and Weekend is not confirmed.

## 3.4 Sufficient sample size for significant result

Supposing the probabilities $p_{i+}$, $p_{+j}$, $p_{ij}$ are constant and sample size $n$ is varying, we can get $\chi^2$-statistics from

$$nX^2 = \sum_{k=1}^{K}\sum_{l=1}^{L} \frac{(p_{kl} - p_{k+}p_{+l})^2}{p_{k+}p_{+l}} \overset{n\to\infty}{\Rightarrow} \chi^2((K-1)(L-1))$$

We know $X^2$ and $L$ for pairs `channel–timegroup` and `channel–weekday`, so, we can get sufficient $K$ for significant results.

For the pair `channel–timegroup` and confidence level 0.95 we have the following equation:

$$n \cdot 0.04624 = \left(\chi^2_{(K-1)(L-1)}\right)^{-1}(0.95) = \left(\chi^2_{6\cdot 3}\right)^{-1}(0.95) = 28.8693$$

$$n_{0.95} \approx 625$$

Table 13: Quetelet relative index table over `channel` and `timegroup`

| | (7.28,189] | (189,370] | (370,550] | (550,732] |
|---|---|---|---|---|
| 0 | -3.41 | -11.26 | 7.43 | 8.87 |
| 1 | -20.79 | -28.94 | 8.92 | **47.23** |
| 2 | 14.67 | **9.57** | -13.38 | -15.12 |
| 3 | -17.81 | -0.97 | 9.12 | 13.98 |
| 4 | -52.33 | 4.91 | **21.76** | 37.72 |
| 5 | -24.86 | 1.73 | 16.78 | 12.36 |
| 6 | **44.90** | 5.00 | -23.86 | -37.08 |

Table 14: Quetelet relative index table over `channel` and `weekday`

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 0 | -9.00 | -6.21 | -9.57 | -3.54 | 9.54 | 22.97 | 31.32 |
| 1 | -8.49 | -2.00 | -7.41 | -10.17 | -10.58 | **42.20** | **54.71** |
| 2 | **9.93** | 3.83 | -5.11 | -8.05 | -5.31 | -10.97 | 20.84 |
| 3 | 0.64 | -4.00 | **18.60** | 7.08 | 0.85 | -44.03 | -21.71 |
| 4 | -10.26 | 6.68 | -5.79 | **8.35** | 1.67 | 21.26 | -20.51 |
| 5 | 2.61 | **8.93** | 2.28 | 3.19 | -10.24 | 9.69 | -30.03 |
| 6 | 0.33 | -4.75 | -2.17 | 0.93 | **7.98** | 2.43 | -3.53 |

| | (7.28,189] | (189,370] | (370,550] | (550,732] | Sum |
|---|---|---|---|---|---|
| 0 | -0.007 (0.00005) | -0.02161 (0.0004) | 0.01362 (0.0002) | 0.01659 (0.0003) | (0.00098) |
| 1 | -0.02558 (0.0006) | -0.03310 (0.001) | 0.00975 (0.0001) | 0.05266 (0.003) | (0.0046) |
| 2 | 0.03280 (0.001) | 0.01989 (0.0004) | -0.02655 (0.0007) | -0.03061 (0.0009) | (0.003) |
| 3 | -0.03889 (0.0015) | -0.00198 (0) | 0.01767 (0.0003) | 0.02765 (0.0008) | (0.0026) |
| 4 | -0.06821 (0.0047) | 0.00595 (0.00004) | 0.02518 (0.0006) | 0.04453 (0.002) | (0.0073) |
| 5 | -0.05743 (0.0033) | 0.00371 (0.00001) | 0.03441 (0.0012) | 0.02587 (0.0006) | (0.005) |
| 6 | 0.11197 (0.013) | 0.01158 (0.00013) | -0.05281 (0.00279) | -0.08375 (0.007) | (0.0225) |
| Sum | (0.0237) | (0.002) | (0.006) | (0.0144) | (0.04624) |

Table 15: $\chi^2$–summary Quetelet index over `channel` and `timegroup`

By similar arguments we obtain sufficient number of observation for confidence level 0.99:

$$n \cdot 0.04624 = \left(\chi^2_{(K-1)(L-1)}\right)^{-1}(0.99) = \left(\chi^2_{6\cdot 3}\right)^{-1}(0.99) = 34.80531$$

$$n_{0.99} \approx 753$$

For the second pair `channel`–`weekday` we know $K = 7, L = 7$. As in the previous case, we solve two equations:

$$n \cdot 0.0124 = \left(\chi^2_{(K-1)(L-1)}\right)^{-1}(0.95) = \left(\chi^2_{6\cdot 6}\right)^{-1}(0.95) = 50.99846$$

$$n_{0.95} \approx 4113$$

By similar arguments we obtain sufficient number of observation for confidence level 0.99:

$$n \cdot 0.0124 = \left(\chi^2_{(K-1)(L-1)}\right)^{-1}(0.99) = \left(\chi^2_{6\cdot 6}\right)^{-1}(0.99) = 58.61921$$

$$n_{0.99} \approx 4727.356$$

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Sum |
|---|---|---|---|---|---|---|---|---|
| 0 | -0.014 (0) | -0.01 (0) | -0.016 (0) | -0.006 (0) | 0.014 (0) | 0.022 (0) | 0.032 (0.001) | (0.002) |
| 1 | -0.008 (0) | -0.002 (0) | -0.007 (0) | -0.01 (0) | -0.009 (0) | 0.024 (0.001) | 0.034 (0.001) | (0.002) |
| 2 | 0.017 (0) | 0.007 (0) | -0.009 (0) | -0.014 (0) | -0.008 (0) | -0.011 (0) | 0.023 (0.001) | (0.001) |
| 3 | 0.001 (0) | -0.007 (0) | 0.033 (0.001) | 0.012 (0) | 0.001 (0) | -0.044 (0.002) | -0.024 (0.001) | (0.004) |
| 4 | -0.01 (0) | 0.007 (0) | -0.006 (0) | 0.009 (0) | 0.002 (0) | 0.013 (0) | -0.013 (0) | (0.001) |
| 5 | 0.005 (0) | 0.017 (0) | 0.004 (0) | 0.006 (0) | -0.017 (0) | 0.01 (0) | -0.035 (0.001) | (0.002) |
| 6 | 0.001 (0) | -0.009 (0) | -0.004 (0) | 0.002 (0) | 0.014 (0) | 0.003 (0) | -0.004 (0) | (0) |
| Sum | (0.001) | (0.001) | (0.002) | (0.001) | (0.001) | (0.003) | (0.005) | (0.0124) |

Table 16: $\chi^2$–summary Quetelet index over `channel` and `weekday`