

ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ

ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

“ВЫСШАЯ ШКОЛА ЭКОНОМИКИ”

АНАЛИЗ СТРАХОВ МОЛОДЕЖИ

ANALYSIS OF FEARS OF YOUNG PEOPLE

Homework Project “2018/2019”

Команда «Fears»

Галиулин Владислав Рафаилович

Лыкова Александра Александровна

Учебная группа мНОД18

1 курс магистратуры

Программа: «Науки о данных»

Факультет компьютерных наук

“Fears” team:

Vladislav Galiulin

Aleksandra Lykova

Group DS

MSc Program “Data Science”

1st Year

Moscow 2018

TABLE OF CONTENTS

1. CHOICE OF THE DATASET (HOMEWORK 1)
2. CLUSTER ANALYSIS (HOMEWORK 2, I)
3. BOOTSTRAP (HOMEWORK 2, II)
4. CONTINGENCY TABLE AND ITS ANALYSIS (HOMEWORK 3)
5. PCA (HOMEWORK 4)
6. LINEAR REGRESSION AND CORRELATION (HOMEWORK 5)

ОБЪЯСНЕНИЕ ВЫБОРА НАБОРА ДАННЫХ.

В рамках данной работы авторами будет сделана попытка выявления неоднородности распространенности специфических фобий среди групп обладающих различными демографическими характеристиками. В ранних работах уже были сделаны попытки оценки различий в фобиях в разных гендерных и возрастных фобиях. Так, было показано, что женщины по сравнению с мужчинами давали более высокие оценки страха для всех объектов и ситуаций. Страхи и фобии неодушевленных предметов чаще встречались у пожилых людей, чем у молодых. Животные страхи были более интенсивными у молодых людей, чем у пожилых. Страх перед полетами усиливается в зависимости от возраста у женщин, но не у мужчин. Таким образом, специфические страхи и фобии неоднородны по половому и возрастному распределению (Fredrikson, 1996).¹

Данные представляют собой результаты опроса молодых людей в возрасте от 15 до 30 лет, проведенного в 2013 году. Исходный набор данных включает в себя 1010 строк (одна строка соответствует набору атрибутов по одному студенту) и 150 атрибутов по различным тематикам, но не по всем собран полный атрибутивный состав информации, полные данные есть по 686 респондентам. Из 150 доступных атрибутов были выбраны 20, которые относятся к информационным блокам 1) Демографические характеристики 2) Фобии респондентов. Данные получены с платформы kaggle².

С помощью анализа этих данных, можно делать выводы о возрастных/гендерных и т.п. различиях в отношении страхов людей. Результаты могут быть использованы в любой индустрии перформанса, например, в киноиндустрии или рекламе - знания своего целевого сегмента и их демографических характеристик позволит отнести каждого клиента к

¹ Fredrikson M. et al. Gender and age differences in the prevalence of specific fears and phobias //Behaviour research and therapy. – 1996. – Т. 34. – №. 1. – С. 33-39.

² <https://www.kaggle.com/miroslavsabo/young-people-survey/home>

определенному кластеру и использовать их фобии как инструмент управления поведением/вниманием аудитории.

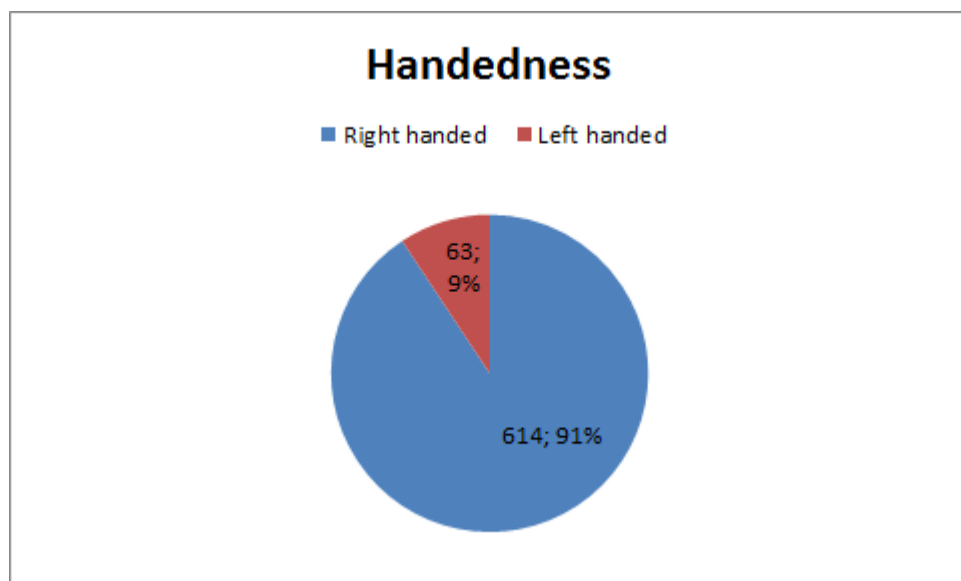
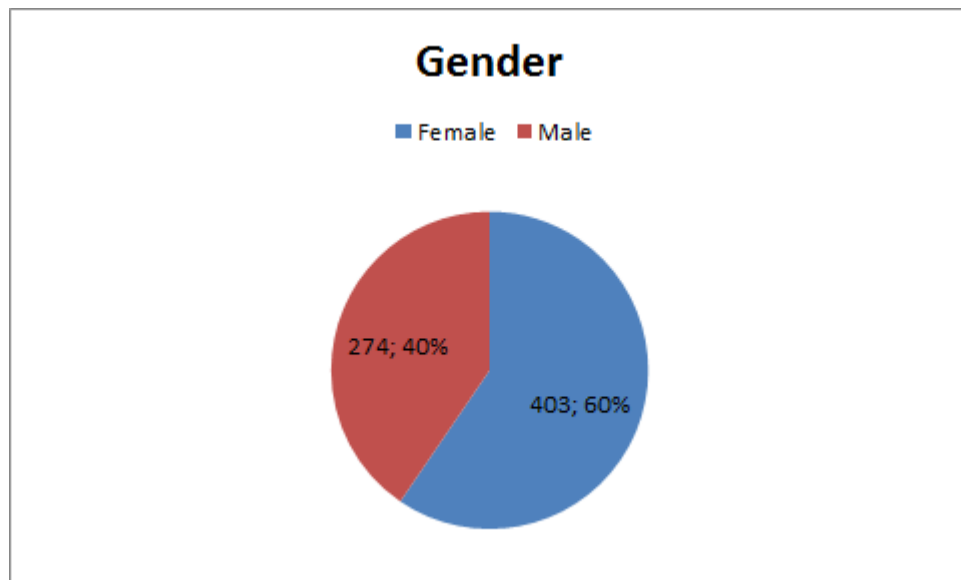
В категории Фобии респондентам было предложено оценить 10 фобий по 5-балльной шкале от 1 (Not afraid at all) до 5 (Very afraid of). Все переменные являются целочисленными.

	Min	Median	Mean	Max
Flying:	1	2	1.994	5
Thunder, lightning:	1	2	2.254	5
Darkness:	1	2	2.254	5
Heights:	1	2	2.573	5
Spiders:	1	3	2.846	5
Snakes:	1	3	3.015	5
Rats, mice:	1	2	2.396	5
Ageing:	1	2	2.529	5
Dangerous dogs:	1	3	3.004	5
Public speaking:	1	3	2.812	5

В категории демографических вопросов респонденты также ответили на 10 вопросов. В результате получились 4 целочисленные переменные и 6 категориальных.

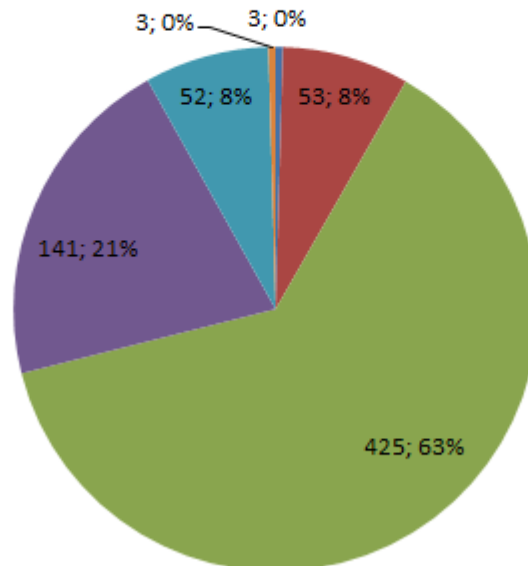
	Min	Median	Mean	Max
Age	15	20	20.35	30

Height	152	172	173.5	203
Weight	41	63	66.15	150
Number of siblings	1	1	1.307	10



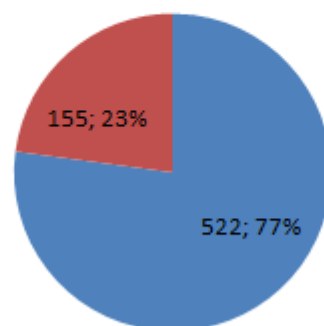
Highest education achieved

- Currently a Primary school pupil
- Primary school
- Secondary school
- College/Bachelor degree
- Masters degree
- Doctorate degree



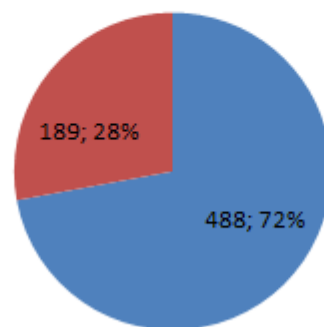
I am the only child

- No
- Yes



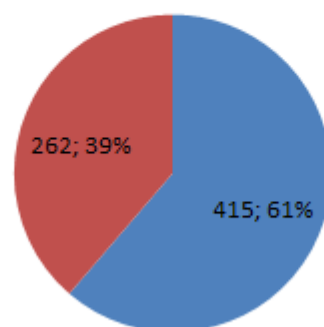
I spent most of my childhood in a

■ City ■ Village



I lived most of my childhood in a:

■ Block of flats house ■ Bungalow



CLUSTER ANALYSIS (HOMEWORK 2, I)

For applying cluster analysis we chose six features, three of them describes demographic characteristics of respondent namely Age, Height and Weight and another three are about fears of these respondents. For clustering we selected such fears to examine as fears of darkness, flying and Rats. We chose intelligent K-Means method, which is implemented as follows:

1. Standardization data by centering and, if needed, normalization;
2. Iteratively finding all Anomalous clusters;
3. Choice the largest K among them or, if K is difficult to specify, setting threshold on the minimum cardinality of a cluster;
4. Application K-Means initialized at chosen Anomalous Cluster centers.

Intelligent K-Means resulted in 5 clusters

Data normalized by range

Anomalous pattern cardinality to discard = 30

Features involved:

Age Mean = 20.35 (On average respondents are at their twenties)

Height Mean = 173.45 (On average respondents have height more than 170cm)

Weight Mean = 66.15 (On average respondents have weight about 66)

Flying Mean = 1.99 (On average respondents are not afraid of flying)

Darkness Mean = 2.25 (On average respondents are not very afraid of darkness)

Rats Mean = 2.40 (On average respondents are not very afraid of rats)

Cluster 1 (111):

Cluster centroid (real) 20.44 169.25 59.81 2.40 3.94 4.16

Cluster centroid (stand) 0.006 -0.082 -0.058 0.101 0.421 0.442

Centroid (% over/under grand mean) 0.0 -2.0 -10.0 20.0 75.0 74.0

Cluster contribution (proper and cumulative) 0.168 0.168

Features significantly larger than average: Darkness (75%), Rats (74%),

Features significantly smaller than average: -

Conclusion: Respondents from this cluster are not very different from other clusters in terms of demos characteristics, but they are more afraid of darkness

and rats than other people from the survey. So please don't take these people to walk in night NY city (everybody knows that there are many rats there at night)

Cluster 2 (228):

Cluster centroid (real) 20.50 176.24 69.74 1.23 1.40 1.32

Cluster centroid (stand) 0.010 0.055 0.033 -0.192 -0.213 -0.27

Centroid (% over/under grand mean) 1.0 2.0 5.0 -38.0 -38.0 -45.0

Cluster contribution (proper and cumulative) 0.139 0.307

Features significantly larger than average: -

Features significantly smaller than average: Flying (-38%), Darkness (-38%), Rats (-45%)

Conclusion: Respondents from this cluster are not very different from other clusters in terms of demos characteristics, but they are less afraid of all fears we analyzed. We concluded that they are fearless and you can offer them any adventure you can imagine. This cluster is the largest among all others. The cluster includes 33.6% of all respondents.

Cluster 3 (92):

Cluster centroid (real) 19.49 170.46 61.83 1.97 3.60 1.33

Cluster centroid (stand) -0.057 -0.059 -0.040 -0.007 0.336 -0.267

Centroid (% over/under grand mean) -4.0 -2.0 -7.0 -1.0 60.0 -45.0

Cluster contribution (proper and cumulative) 0.068 0.375

Features significantly larger than average: Darkness (60%)

Features significantly smaller than average: Rats (-45%)

Conclusion: Respondents from this cluster are not very different from other clusters in terms of demos characteristics, but they are less afraid of rats and more afraid of darkness. Please don't turn off the light when these people fall asleep.

Cluster 4 (120):

Cluster centroid (real) 20.58 173.19 67.20 3.68 1.81 2.23

Cluster centroid (stand) 0.016 -0.005 0.010 0.420 -0.111 -0.041

Centroid (% over/under grand mean) 1.0 -0.0 2.0 84.0 -20.0 -7.0

Cluster contribution (proper and cumulative) 0.088 0.464

Features significantly larger than average: Flying (84%)

Features significantly smaller than average: -

Conclusion: Respondents from this cluster are not very different from other clusters in terms of demos characteristics, but they are much more afraid of flying than the average respondent. Please buy them train ticket.

Cluster 5 (126):

Cluster centroid (real) 20.40 174.56 67.38 1.44 1.75 3.73

Cluster centroid (stand) 0.003 0.022 0.011 -0.137 -0.125 0.334

Centroid (% over/under grand mean) 0.0 1.0 2.0 -28.0 -22.0 56.0

Cluster contribution (proper and cumulative) 0.071 0.534

Features significantly larger than average: Rats (56%)

Features significantly smaller than average: -

Conclusion: Respondents from this cluster are not very different from other clusters in terms of demos characteristics, but they are more afraid of rats. Do not show them the cartoon "Ratatouille".

Intelligent K-Means resulted in 9 clusters

Data normalized by range

Anomalous pattern cardinality to discard = 5

Features involved:

Age Mean = 20.35 (On average respondents are at their twenties)

Height Mean = 173.45 (On average respondents have height more than 170cm)

Weight Mean = 66.15 (On average respondents have weight about 66)

Flying Mean = 1.99 (On average respondents are not afraid of flying)

Darkness Mean = 2.25 (On average respondents are not very afraid of darkness)

Rats Mean = 2.40 (On average respondents are not very afraid of rats)

Cluster 1 (51):

Cluster centroid (real) 20.51 168.49 60.43 3.43 3.65 4.49

Cluster centroid (stand) 0.011 -0.097 -0.052 0.359 0.348 0.524

Centroid (% over/under grand mean) 1.0 -3.0 -9.0 72.0 62.0 87.0

Cluster contribution (proper and cumulative) 0.105 0.105

Features significantly larger than average: Flying (72%), Darkness (62%), Rats (87%)

Features significantly smaller than average: -

Conclusion: Respondents from this cluster are not very different from other

clusters in terms of demos characteristics, but they are much more afraid of all (flying, darkness and rats) than the average respondent.

Cluster 2 (119):

Cluster centroid (real) 21.17 181.78 76.64 1.08 1.24 1.16

Cluster centroid (stand) 0.055 0.163 0.096 -0.230 -0.253 -0.30

Centroid (% over/under grand mean) 4.0 5.0 16.0 -46.0 -45.0 -52.0

Cluster contribution (proper and cumulative) 0.115 0.220

Features significantly larger than average: -

Features significantly smaller than average: Flying (-46%), Darkness (-45%), Rats (-52%)

Conclusion: Respondents from this cluster are not very different from other clusters in terms of demos characteristics, but they are less afraid of all (flying, darkness and rats) than the average respondent.

Cluster 3 (64):

Cluster centroid (real) 19.45 168.94 60.08 2.16 3.84 1.34

Cluster centroid (stand) -0.060 -0.089 -0.056 0.041 0.397 -0.263

Centroid (% over/under grand mean) -4.0 -3.0 -9.0 8.0 71.0 -44.0

Cluster contribution (proper and cumulative) 0.060 0.280

Features significantly larger than average: Darkness (71%)

Features significantly smaller than average: Rats (-44%)

Conclusion: Respondents from this cluster are not very different from other clusters in terms of demos characteristics. They are less afraid of rats, but they are much more afraid of darkness than the average respondent.

Cluster 4 (53):

Cluster centroid (real) 19.66 168.72 60.62 3.85 1.45 2.68

Cluster centroid (stand) -0.046 -0.093 -0.051 0.464 -0.200 0.071

Centroid (% over/under grand mean) -3.0 -3.0 -8.0 93.0 -36.0 12.0

Cluster contribution (proper and cumulative) 0.056 0.336

Features significantly larger than average: Flying (93%)

Features significantly smaller than average: Darkness (-36%)

Conclusion: Respondents from this cluster are not very different from other clusters in terms of demos characteristics. They are less afraid of darkness, but they are much more afraid of flying than the average respondent.

Cluster 5 (96):

Cluster centroid (real) 20.67 173.89 66.38 1.39 1.54 3.81

Cluster centroid (stand) 0.021 0.008 0.002 -0.152 -0.178 0.354

Centroid (% over/under grand mean) 2.0 0.0 0.0 -31.0 -32.0 59.0

Cluster contribution (proper and cumulative) 0.067 0.403

Features significantly larger than average: Rats (59%)

Features significantly smaller than average: Flying (-31%), Darkness (-32%)

Conclusion: Respondents from this cluster are not very different from other clusters in terms of demos characteristics. They are less afraid of flying and darkness, but they are much more afraid of rats than the average respondent.

Cluster 6 (70):

Cluster centroid (real) 19.94 169.69 59.83 1.30 3.84 3.94

Cluster centroid (stand) -0.027 -0.074 -0.058 -0.174 0.397 0.387

Centroid (% over/under grand mean) -2.0 -2.0 -10.0 -35.0 70.0 65.0

Cluster contribution (proper and cumulative) 0.093 0.496

Features significantly larger than average: Darkness (70%), Rats (65%)

Features significantly smaller than average: Flying (-35%)

Conclusion: Respondents from this cluster are not very different from other clusters in terms of demos characteristics. They are less afraid of flying, but they are much more afraid of darkness and rats than the average respondent.

Cluster 7 (123):

Cluster centroid (real) 19.54 167.56 58.96 1.41 1.79 1.35

Cluster centroid (stand) -0.054 -0.116 -0.066 -0.147 -0.116 -0.262

Centroid (% over/under grand mean) -4.0 -3.0 -11.0 -29.0 -21.0 -44.0

Cluster contribution (proper and cumulative) 0.059 0.555

Features significantly larger than average: -

Features significantly smaller than average: Rats (-44%)

Conclusion: Respondents from this cluster are not very different from other clusters in terms of demos characteristics. They are less afraid of rats than the average respondent.

Cluster 8 (68):

Cluster centroid (real) 20.44 185.47 81.38 2.53 1.74 1.96

Cluster centroid (stand) 0.006 0.236 0.140 0.134 -0.130 -0.110

Centroid (% over/under grand mean) 0.0 7.0 23.0 27.0 -23.0 -18.0

Cluster contribution (proper and cumulative) 0.032 0.587

Features significantly larger than average: -

Features significantly smaller than average: -

Conclusion: there is no significant difference. Respondents from this cluster are about representative respondents of all dataset.

Cluster 9 (33):

Cluster centroid (real) 22.76 171.42 65.91 4.12 3.45 2.61

Cluster centroid (stand) 0.161 -0.040 -0.002 0.532 0.300 0.053

Centroid (% over/under grand mean) 12.0 -1.0 -0.0 107.0 53.0 9.0

Cluster contribution (proper and cumulative) 0.051 0.638

Features significantly larger than average: Flying (107%), Darkness (53%)

Features significantly smaller than average: -

Conclusion: Respondents from this cluster are not very different from other clusters in terms of demos characteristics, but they are much more afraid of flying and darkness than the average respondent.

It can be noted that obtaining clusters for both partitions do not have statistically significant differences in demographic characteristics. Then compare the partitioning of clusters on the characteristics of fear levels.

	K=5				K=9			
	% of entities	Flying	Darkness	Rats	% of entities	Flying	Darkness	Rats
Cluster 1	16.40%	-	75%	74%	7.53%	72%	62%	87%
Cluster 2	33.68%	-38%	-38%	-45%	17.58%	-46%	-45%	-52%
Cluster 3	13.59%	-	60%	-45%	9.45%	-	71%	-44%
Cluster 4	17.73%	84%	-	-	7.83%	93%	-36%	-
Cluster 5	18.61%	-	-	56%	14.18%	-31%	-32%	59%
Cluster 6	-	-	-	-	10.34%	-35%	70%	65%
Cluster 7	-	-	-	-	18.17%	-	-	-44%
Cluster 8	-	-	-	-	10.04%	-	-	-
Cluster 9	-	-	-	-	4.87%	107%	53%	-
Cumulative cluster contribution	0.534				0.638			

When splitting into 9 clusters, we lose a smaller amount of information, but at the same time we get several clusters that make up less than 10% of the entire sample. For further analysis, we decided to take a partition into 5 clusters.



BOOTSTRAP (HOMEWORK 2, II)

For further research we chose 5 cluster partition and among these clusters we applied bootstrap over the first and second clusters (you can see them below). Such **qualitative feature as** “Rats” (see description above) was used in bootstrap.

Cluster 1 (111):

Cluster centroid (real)	20.44	169.25	59.81	2.40	3.94	4.16
Cluster centroid (stand)	0.006	-0.082	-0.058	0.101	0.421	0.442
Centroid (% over/under grand mean)	0.0	-2.0	-10.0	20.0	75.0	74.0
Cluster contribution (proper and cumulative)	0.168		0.168			

Features significantly larger than average: **Darkness (75%), Rats (74%),**

Features significantly smaller than average:

Cluster 2 (228):

Cluster centroid (real)	20.50	176.24	69.74	1.23	1.40	1.32
Cluster centroid (stand)	0.010	0.055	0.033	-0.192	-0.213	-0.27
Centroid (% over/under grand mean)	1.0	2.0	5.0	-38.0	-38.0	-45.0
Cluster contribution (proper and cumulative)	0.139		0.307			

Features significantly larger than average:

Features significantly smaller than average: **Flying (-38%), Darkness (-38%), Rats (-45%)**

2.1 Comparison of the feature “Rats” between two clusters with using bootstrap

Comparison procedure³:

- 1) Bootstrap distributions of 5000 trial means in Cluster1 and in Cluster2
- 2) Compute the difference $D = m_1 - m_2$, where m_1 all trial means from cluster1 and m_2 is all trial means from cluster2. Visualisation of computed difference is presented below (graph 1)
- 3) Find 95% confidence interval
 - a) Pivotal method - find the mean and standard deviation of difference D , compute confidence intervals with 3-sigmas rule. We found that the mean of D equals to 2.846 and standard deviation equals to 0.079,

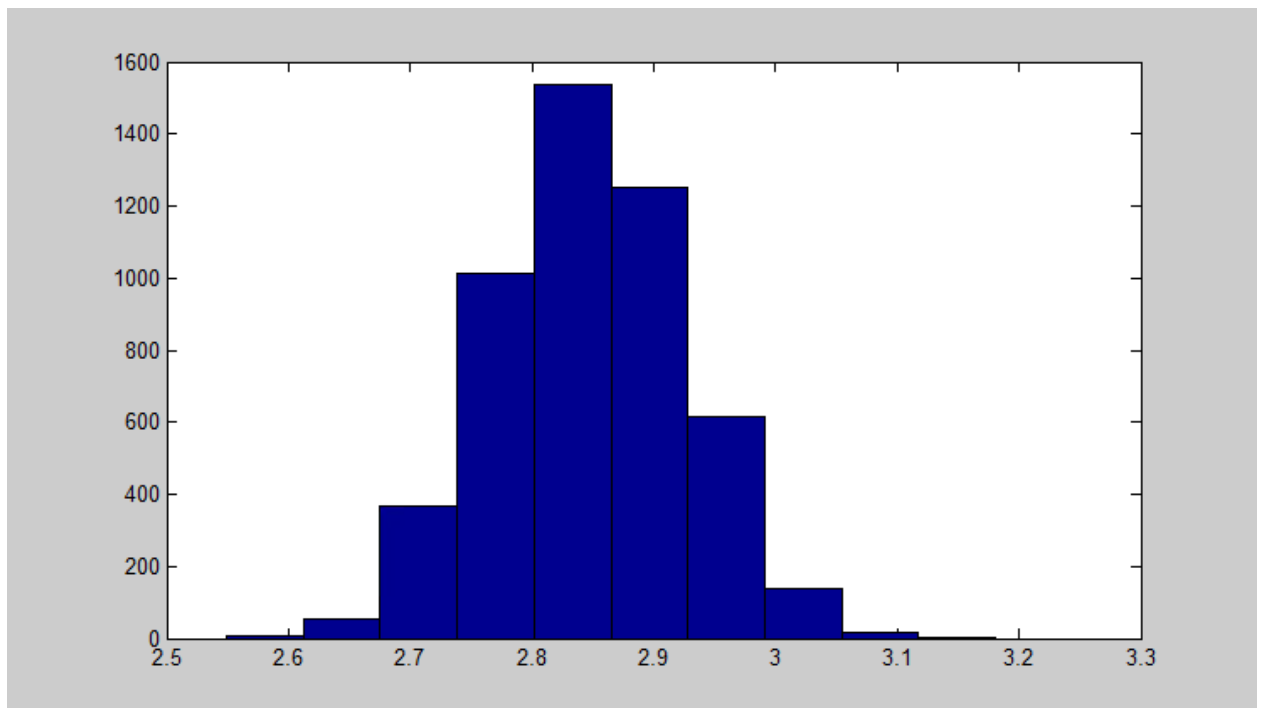
³ All calculations are performed in MATLAB software, the code used is presented in applications

so the left border is equal to $2.846 - 3 \cdot 0.079 = \mathbf{2.692}$ and right border is equal to $2.846 + 3 \cdot 0.079 = \mathbf{3.001}$

b) Non - pivotal method - arrange the observations in ascending order, then remove 2.5% of the observations at each end. So the borders are **2.694 (left) and 3.005 (right)**.

4) Check whether zero is in the intervals we found. As we can see 0 does not belong to our intervals in both methods (pivotal and non-pivotal)

5) **To make conclusion: with 95% confidence we can conclude that means of Rats feature in clusters one and two are not equal. Moreover, we can say that the mean of Rats variable from cluster 1 is greater than the mean of Rats variable from cluster 2. The results obtained correspond to the results of cluster analysis**



Graph 1. Distribution of difference (D) between trial means of Rats feature in clusters 1 and 2

2.2 Take a feature, find the 95% confidence interval for its grand mean by using bootstrap

Finding confidence interval procedure⁴:

- 1) Bootstrap distributions of 5000 trial means of qualitative feature “rats”
- 2) Find 95% confidence interval
 - c) Pivotal method - find the mean and standard deviation of trial means (mr), compute confidence intervals with 3-sigmas rule. We found that the mean of mr equals to 2.396 and standard deviation equals to 0.052,
so the left border is equal to $2.396 - 3 * 0.052 = \mathbf{2.294}$ and right border is equal to $2.396 + 3 * 0.052 = \mathbf{2.499}$
 - d) Non - pivotal method - arrange the observations in ascending order, then remove 2.5% of the observations at each end. So the borders are **2.294 (left) and 2.501 (right)**.

2.3 Take a cluster, and compare the grand mean with the within-cluster mean for the feature by using bootstrap

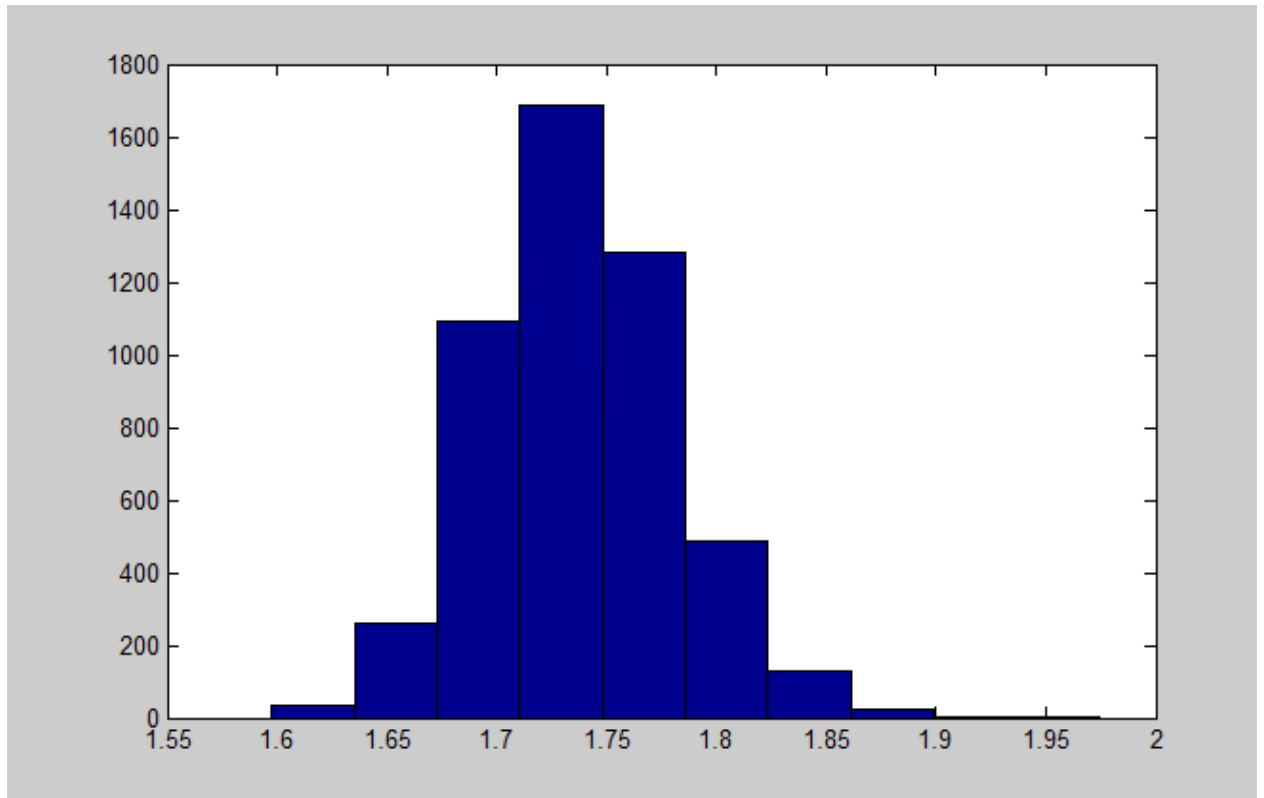
Comparison procedure⁵:

- 1) Bootstrap distributions of 5000 trial means for Cluster1 and Grand Mean
- 2) Compute the quotients $Q = m1 / mr$, where m1 all trial means from cluster1 and mr is all trial means for grand mean. Visualisation of computed quotients is presented below (Graph 2)
- 3) Find 95% confidence interval
 - e) Pivotal method - find the mean and standard deviation of quotients Q, compute confidence intervals with 3-sigmas rule. We found that the mean of Q equals to 1.7373 and standard deviation equals to 0.043,
so the left border is equal to $1.7373 - 3 * 0.043 = \mathbf{1.652}$ and right border is equal to $1.7373 + 3 * 0.043 = \mathbf{1.822}$
 - f) Non - pivotal method - arrange the observations in ascending order, then remove 2.5% of the observations at each end. So the borders are **1.659 (left) and 1.827 (right)**.
- 4) Check whether 1 is in the intervals we found. As we can see 1 does not belong to our intervals in both methods (pivotal and non-pivotal)

⁴ All calculations are performed in MATLAB software, the code used is presented in applications

⁵ All calculations are performed in MATLAB software, the code used is presented in applications

5) To make conclusion: with 95% confidence we can conclude that means of Rats feature in clusters one and in overall dataset are not equal. Moreover, we can say that the mean of Rats variable from cluster 1 is greater than the overall mean of Rats variable in dataset. The results obtained correspond to the results of cluster analysis.



Graph 2. Distribution of quotients (Q) between trial means of Rats feature in clusters 1 and grand mean

CONTINGENCY TABLE (HOMEWORK 3)

3.1 Consider three nominal features (one of them, not more, may be taken from nominal features in your data)

We chose three features out of our dataset to carry out the analysis using contingency tables. The first one was **Education** variable and it was initially set as a nominal variable with following six levels: college/bachelor degree; currently a primary school pupil; doctorate degree; masters degree; primary school; secondary school. Other two variables were two other variables (**Age** and **Rats Fear**) have been reduced to the form categories. We used 33% and 67% quantile constraints to convert quantitative variables into categories. Thus, the categories obtained by us can be interpreted as Low/Medium / High levels of the variable value.

3.2 Build two contingency tables over them: present a conditional frequency table and Quetelet relative index tables. Make comments on relations between categories of the common (to both tables) feature and two others.

The first table presented below is the contingency table⁶ for two variables Education and categorized Age (Table 1) - every cell at the intersection of columns and rows contains a number that equals to the **quantity of respondents** that simultaneously falling under the categories of both variables.

Table 1

quantity of respondents Education Level	Age Category			
	Low Aged	Middle Aged	High Aged	Total
secondary school	204	163	58	425
college/bachelor degree	38	48	55	141
primary school	51	1	1	53
masters degree	1	4	47	52
doctorate degree			3	3
currently a primary school pupil	3			3
Total	297	216	164	677

⁶ All calculations for this section was performed in online service Google Tables

For example, in Table 1 we can see that among 677 respondents 48 are Middle Aged and have college/bachelor education level simultaneously, secondary school education level is major class over all age levels.

The second table presented is also the contingency table for two variables Education and categorized Age (Table 2) - every cell at the intersection of columns and rows contains an number that equals to the **proportion of respondents in percents** that simultaneously falling under the categories of both variables. For example, in Table 2 we can see that among all respondents 30,13% are in a Low Age category and have secondary school education level.

Table 2. Relative frequency Education Level vs Age Category

quantity of respondents Education Level	Age Category			
	Low Aged	Middle Aged	High Aged	Total
secondary school	30,13%	24,08%	8,57%	62,78%
college/bachelor degree	5,61%	7,09%	8,12%	20,83%
primary school	7,53%	0,15%	0,15%	7,83%
masters degree	0,15%	0,59%	6,94%	7,68%
doctorate degree			0,44%	0,44%
currently a primary school pupil	0,44%			0,44%
Total	43,87%	31,91%	24,22%	100,00%

The third table presented illustrates conditional frequency for two variables Education and categorized Age (Table 3) - every cell represents the **proportion of respondents with particular education level among the group of age category**.

For example, in Table 3 we can see that among all high aged respondents only 7,68% have masters degree.

Table 3. Conditional frequency Education Level in Age Category

quantity of respondents Education Level	Age Category			
	Low Aged	Middle Aged	High Aged	Total
secondary school	68,69%	75,46%	35,37%	62,78%
college/bachelor degree	12,79%	22,22%	33,54%	20,83%
primary school	17,17%	0,46%	0,61%	7,83%
masters degree	0,34%	1,85%	28,66%	7,68%
doctorate degree			1,83%	0,44%
currently a primary school pupil	1,01%			0,44%

Total	100,00%	100,00%	100,00%	100,00%
-------	---------	---------	---------	---------

The forth table presented consists of conditional frequency for two variables Education and categorized Age (Table 4) - every cell represents the **proportion of respondents with particular age category among education levels** . For example, in Table 4 we can see that among respondents with primary school education level 96,23% are low aged.

Table 4. Conditional frequency of Age Category in every Education Level

quantity of respondents Education Level	Age Category			
	Low Aged	Middle Aged	High Aged	Total
secondary school	48,00%	38,35%	13,65%	100,00%
college/bachelor degree	26,95%	34,04%	39,01%	100,00%
primary school	96,23%	1,89%	1,89%	100,00%
masters degree	1,92%	7,69%	90,38%	100,00%
doctorate degree			100,00%	100,00%
currently a primary school pupil	100,00%			100,00%
Total	43,87%	31,91%	24,22%	100,00%

However it's hard to conclude is this level conditional or relative frequency high or low. To be able to understand the difference more precisely let us to move on for Quetelet index computation. The (empirical) probability that category I occurs under condition of k can be expressed as $P(I/k) = p_{kI}/p_k$. The probability $P(I)$ of the category I with no condition is just $p_{+I} = N_{+I}/N$. Similar notation is used when I and k are swapped. The relative difference between the two probabilities is referred to as (relative) Quetelet index⁷:

$$q(I/k) = \frac{p(I/k) - p(I)}{p(I)}$$

where $P(I) = N_{+I}/N$, $P(k) = N_{+k}/N$, $P(I/k) = N_{kI}/N_{+k}$. That is, Quetelet index expresses correlation between categories k and I as the relative change in the probability of I when k is taken into account.

Table 5 presents Quetelet index $q(ED_k/AGE_I)$:

⁷ Mirkin B. Core Concepts in Data Analysis: Summarization, Correlation, Visualization //Department of Computer Science and Information Systems, Birkbeck, University of London, Department of Data Analysis and Machine Intelligence, Higher School of Economics, 11 Pokrovski Boulevard, Moscow RF. – 2010.

$$q(EDk/AGEl) = \frac{p(ED/AGEl) - p(ED)}{p(ED)}$$

Table 5. Education/Age category Cross classification Quetelet coefficients, %

<i>quantity of respondents</i>	<i>Age Category</i>			
<i>Education Level</i>	Low Aged	Middle Aged	High Aged	P(ED)
secondary school	9,41%	20,21%	-43,66%	62,78%
college/bachelor degree	-38,57%	6,70%	61,02%	20,83%
primary school	119,34%	-94,09%	-92,21%	7,83%
masters degree	-95,62%	-75,89%	273,11%	7,68%
doctorate degree	-100,00%	-100,00%	312,80%	0,44%
currently a primary school pupil	127,95%	-100,00%	-100,00%	0,44%

From this table we can see that, for example, condition “Age Category=High” raises the frequency of the master degree edu level category by 273.11% or condition “Edu Level=primary school” decreases the frequency of the middle aged respondents by 94.09%



On the next stage we repeated all calculations for another pair of variables Education and categorized Rats Fear (Table 6) - every cell at the intersection of columns and rows contains an number that equals to the **quantity of respondents** that simultaneously falling under the categories of both variables.

Table 6

<i>quantity of respondents</i>	<i>Rats Fear Category</i>			
<i>Education Level</i>	Middle Fear	Low Fear	High Fear	Total
secondary school	171	160	94	425
college/bachelor degree	52	54	35	141
primary school	17	23	13	53
masters degree	20	13	19	52
doctorate degree	1	1	1	3
currently a primary school pupil	1	2		3
Total	262	253	162	677

For example, in Table 6 we can see that among 677 respondents 160 have relatively low fear level and secondary school education level simultaneously.

The next table is also the contingency table for two variables Education and categorized Rats Fear (Table 7) - every cell at the intersection of columns and rows contains an number that equals to the **proportion of respondents in percents** that simultaneously falling under the categories of both variables. For example, in Table 7 we can see that among all respondents 23,63% are in a Low Fear category and have secondary school education level.

Table 7. Relative frequency Education Level vs Age Category

<i>% of respondents</i> <i>Education Level</i>	<i>Rats Fear Category</i>			
	Middle Fear	Low Fear	High Fear	Total
secondary school	25,26%	23,63%	13,88%	62,78%
college/bachelor degree	7,68%	7,98%	5,17%	20,83%
primary school	2,51%	3,40%	1,92%	7,83%
masters degree	2,95%	1,92%	2,81%	7,68%
doctorate degree	0,15%	0,15%	0,15%	0,44%
currently a primary school pupil	0,15%	0,30%		0,44%
Total	38,70%	37,37%	23,93%	100,00%

The table below presents conditional frequency for two variables Education and categorized Rats Fear (Table 8) - every cell represents the **proportion of respondents with particular education level among the group of Fear category**. For example, in Table 8 we can see that among all respondents with high fear level only 11,73% have masters degree.

Table 8. Conditional frequency Education Level in Age Category

<i>quantity of respondents</i> <i>Education Level</i>	<i>Rats Fear Category</i>			
	Middle Fear	Low Fear	High Fear	Total
secondary school	65,27%	63,24%	58,02%	62,78%
college/bachelor degree	19,85%	21,34%	21,60%	20,83%
primary school	6,49%	9,09%	8,02%	7,83%
masters degree	7,63%	5,14%	11,73%	7,68%
doctorate degree	0,38%	0,40%	0,62%	0,44%

currently a primary school pupil	0,38%	0,79%	0,44%
Total	100,00%	100,00%	100,00%

The eighth table presents conditional frequency for two variables Education and categorized Fear (Table 9) - every cell represents the **proportion of respondents with particular rats Fear category among education levels**. For example, in Table 9 we can see that among respondents with primary school education level 43,40% are with low fear level.

Table 9. Conditional frequency of Age Category in every Education Level

quantity of respondents Education Level	Rats Fear Category			
	Middle Fear	Low Fear	High Fear	Total
secondary school	40,24%	37,65%	22,12%	100,00%
college/bachelor degree	36,88%	38,30%	24,82%	100,00%
primary school	32,08%	43,40%	24,53%	100,00%
masters degree	38,46%	25,00%	36,54%	100,00%
doctorate degree	33,33%	33,33%	33,33%	100,00%
currently a primary school pupil	33,33%	66,67%		100,00%
Total	38,70%	37,37%	23,93%	100,00%

Table 10 presents Quetelet index $q(EDk/FearRatsI)$:

$$q(EDk/AGEI) = \frac{p(FearRats/ED) - p(FearRats)}{p(FearRats)}$$

Table10. Education/Rats category Cross classification Quetelet coefficients, %

Education Level	Rats Fear Category		
	Middle Fear	Low Fear	High Fear
secondary school	3,97%	0,74%	-7,57%
college/bachelor degree	-4,70%	2,48%	3,73%
primary school	-17,12%	16,12%	2,50%
masters degree	-0,62%	-33,10%	52,69%
doctorate degree	-13,87%	-10,80%	39,30%
currently a primary school pupil	-13,87%	78,39%	-100,00%
P(RatFear Category)	38,70%	37,37%	23,93%

From this table we can see that, for example, condition “Fear Category=High” raises the frequency of the master degree edu level category by 52.69% or condition “Edu Level=primary school” decreases the frequency of the middle fear respondents by 17.12%

3.3 Compute and visualize the chi-square- summary_Quetelet_index over both tables. Comment on the meaning of the values in the data analysis context.

Summary Quetelet correlation index Q as the sum of pair-wise Quetelet indexes weighted by their frequencies/probabilities⁸:

$$Q = \sum_{k=1}^K \sum_{l=1}^L p_{kl} q(l, k) = \sum_{k=1}^K \sum_{l=1}^L p_{kl} \left(\frac{p_{kl}}{p_{k+} p_{+l}} - 1 \right) = \sum_{k=1}^K \sum_{l=1}^L \frac{p_{kl}^2}{p_{k+} p_{+l}} - 1$$

The right-hand expression for Q is equal to chi-squared correlation coefficient proposed by K. Pearson (1901) as a measure of deviation of the contingency table entries from the statistical independence ($\chi^2 = NQ$).

So Summary Quetelet correlation index Q was calculated for variables Education Level and Age firstly (see Table 5 above). The result we got equals to 813,25%. It means that on average knowledge about Age category ‘adds’ 813,25% to frequency Education Level. Then the same procedure was used to calculate Quetelet correlation index Q for variables Education Level and Rat Fear Category (see Table 10 above). The result we got equals to 65,72%. It means that on average knowledge about Fear category ‘adds’ 65,72% to frequency Education Level⁹.

3.4 Tell what numbers of observations would suffice to see the features as associated at 95% confidence level; 99% confidence level.

Under the hypothesis that the features are independent in the population, and entity sampling has been done randomly and independently, the density

⁸ Mirkin B. Core Concepts in Data Analysis: Summarization, Correlation, Visualization //Department of Computer Science and Information Systems, Birkbeck, University of London, Department of Data Analysis and Machine Intelligence, Higher School of Economics, 11 Pokrovski Boulevard, Moscow RF. – 2010.

⁹ [Fears Contingency Table](#)

function of random variable χ^2 tends to distribution χ^2 with $f=(K-1)(L-1)$ degrees of freedom.

Lets apply this theorem to our case. We have $K=6, L=3$, therefore $f=9$. At $f=9$, there is a 5% chance that the χ^2 value will be greater than 16,92 if the hypothesis of independence is true. In our case for the first situation (Education Level vs Age Category) $\chi^2 = 8,133$. **To see features associated we need $\chi^2 > 16,92$ so that the independence could be rejected, therefore $N \cdot 8,133 > 16,92$ and $N \geq 3$.** Another case, there is a 1% chance that the χ^2 value will be greater than 21,7 if the hypothesis of independence is true. In our case for the first situation (Education Level vs Age Category) $\chi^2 = 8,133$. **To see features associated we need $\chi^2 > 21,7$ so that the independence could be rejected, therefore $N \cdot 8,133 > 21,7$ and $N \geq 3$.**

Lets apply this theorem to the second pair of variables. We have $K=6, L=3$, therefore $f=9$. At $f=9$, there is a 5% chance that the χ^2 value will be greater than 16,92 if the hypothesis of independence is true. In our case for the first situation (Education Level vs Age Category) $\chi^2 = 0,657$. **To see features associated we need $\chi^2 > 16,92$ so that the independence could be rejected, therefore $N \cdot 0,657 > 16,92$ and $N \geq 26$.** Another case, there is a 1% chance that the χ^2 value will be greater than 21,7 if the hypothesis of independence is true. In our case for the first situation (Education Level vs Age Category) $\chi^2 = 0,657$. **To see features associated we need $\chi^2 > 21,7$ so that the independence could be rejected, therefore $N \cdot 0,657 > 21,7$ and $N \geq 33$.**

4. PCA (HOMEWORK 4)

6 variables were chosen for the Principal Component Analysis. All of them characterize the level of fear: Flying, Heights, Spiders, Snakes, Rats, Darkness. Using factor analysis, we want to explore the relationship between the values of variables. It can be assumed that the fear of heights and flights is correlated. So, having both variables in the data leads to data redundancy. The same can be assumed about the fear of spiders, snakes and rats.

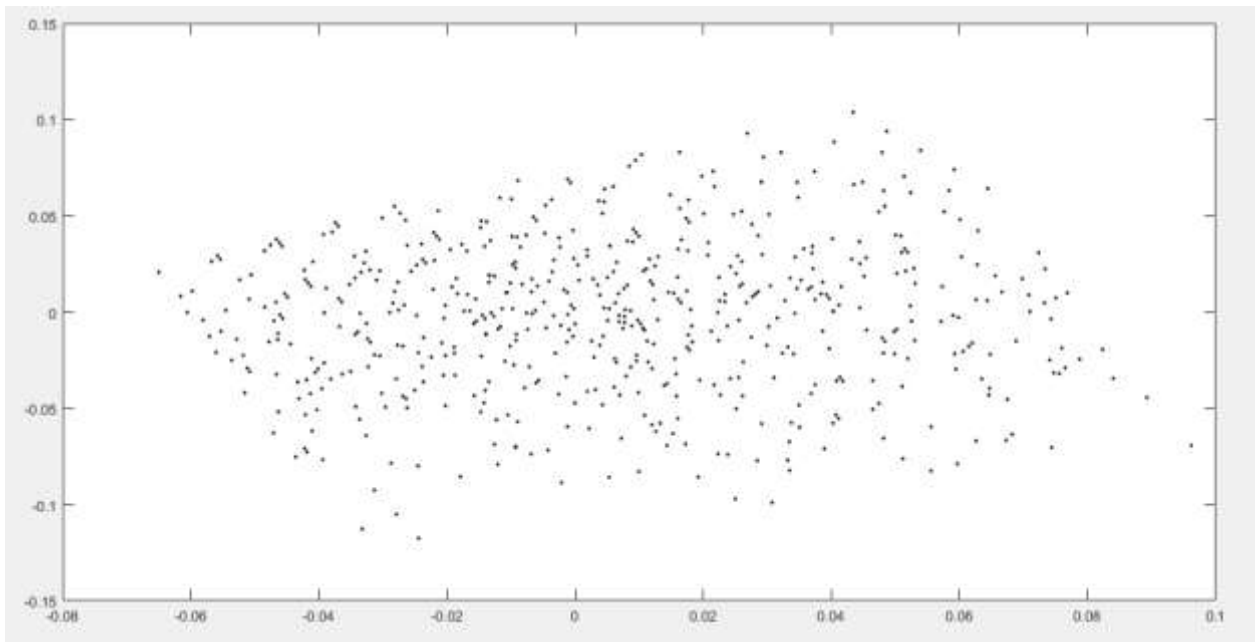
After applying Principal Component Analysis we got the following results:
Data Scatter= 7499,69

	PC1	PC2	PC3	PC4	PC5	PC6
contributions (naturally)	3272.152	1190.817	997.1197	764.4826	703.043	703.043
contributions (per cent)	43.63%	15.88%	13.30%	10.19%	10.19%	7.63%

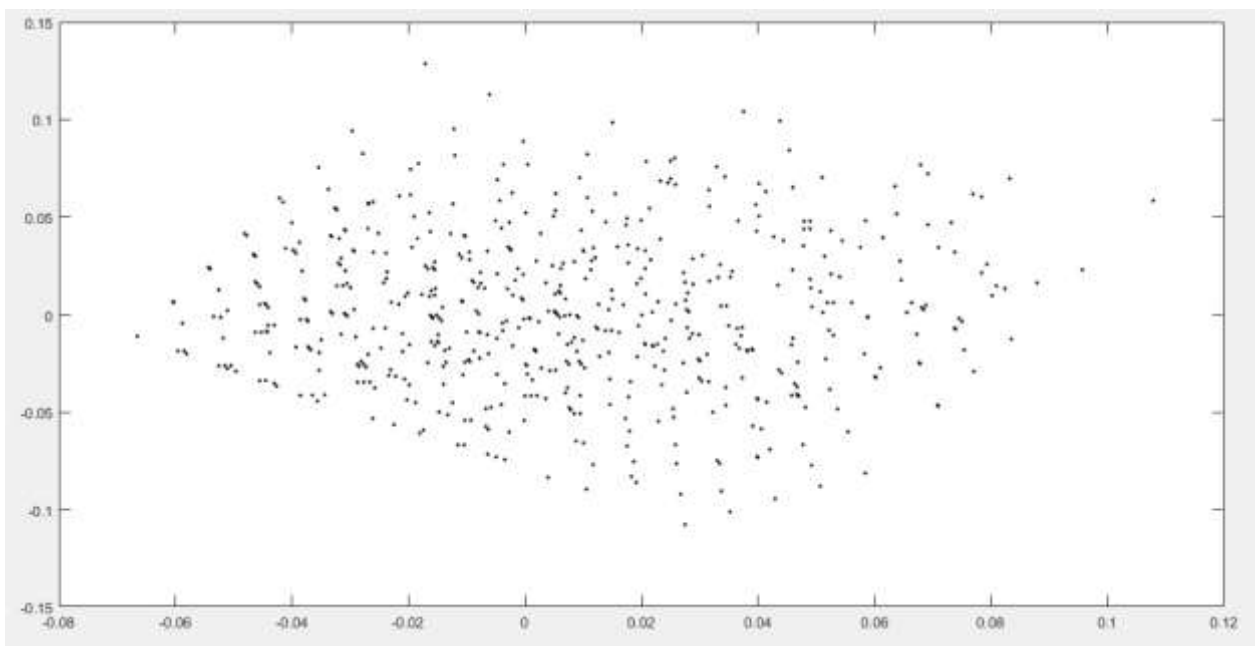
Feature loadings

	Flying	Heights	Spiders	Snakes	Rats	Darkness
PC1	0.29	0.33	0.44	0.49	0.47	0.39
PC2	0.55	0.59	-0.31	-0.30	-0.33	0.22
PC3	-0.65	0.26	0.20	-0.28	-0.16	0.60
PC4	0.38	-0.69	0.06	-0.16	-0.12	0.59
PC5	0.17	0.03	0.79	-0.12	-0.49	-0.29
PC6	0.11	0.02	0.19	-0.74	0.62	-0.14

As we can see the first principal component is responsible for the fear of different living creatures. The second principal component is responsible for the fear of flying and heights. The third principal component is responsible for the fear of darkness and for the fearlessness of flying, etc. And the first three components explain the data at 72.80%.



Visualization the data with these features using standardization
with normalization over ranges



Visualize the data with these features using standardization
with normalization over standard deviations


After this we apply the conventional PCA for the visualizations. And they looked exactly the same as the graphics above. Because of conventional PCA approach leads to the same scoring and loading vectors as the model-based PCA.



Covariance matrix coincides, up to a constant factor, with matrix $A = X' * X$, provided that X is centered. Matrix A is in the core of Singular triplets. Working with eigenvectors of A is equivalent to working with singular vectors of X . Also there is no big difference between 2 types of normalization, because of all of using features have a small range from 2 to 4.

Hidden factors we can interpret as the existence of a small fear of all 6 species, especially snakes and spiders.

Flying	Heights	Spiders	Snakes	Rats	Darkness
0.13	0.17	0.19	0.20	0.16	0.15



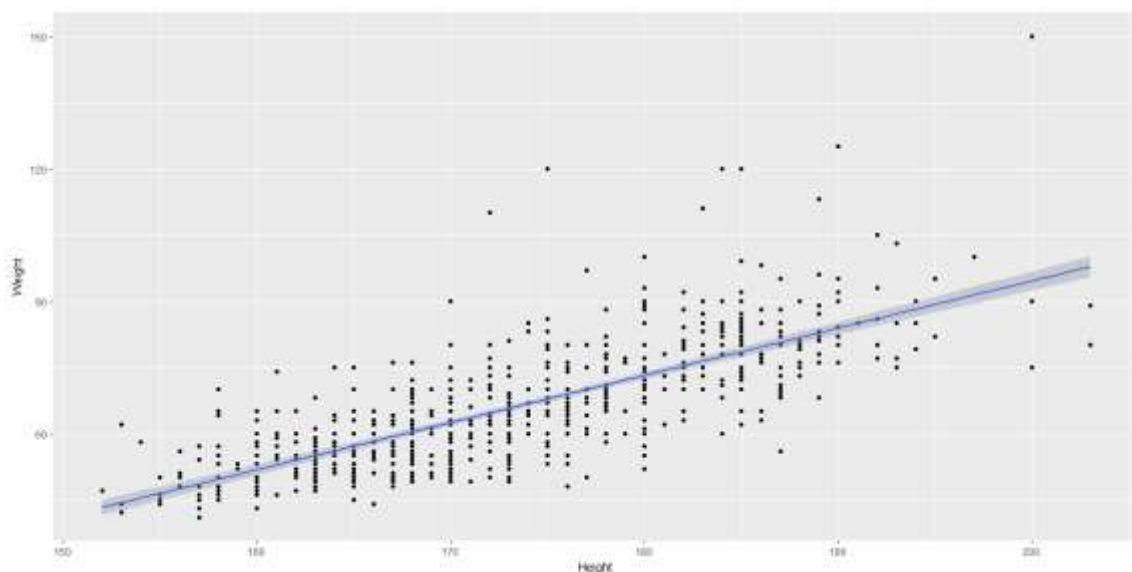
LINEAR REGRESSION AND CORRELATION COEFFICIENT (HOMEWORK 5)

5.1 FIND TWO FEATURES IN YOUR DATASET WITH MORE OR LESS “LINEAR-LIKE” SCATTERPLOT.

Among our set of features we defined two of them that could be more appropriate for regression analysis application so we chose such variables as Height and Weight. Both of the features are presented as quantity measures (you can find descriptive statistics of our measures in the first section of current work).

5.2 DISPLAY THE SCATTER-PLOT.

Below you can find the scatter-plot for features Height (x-axis) and Weight (y-axis), we added the blue line to the plot to illustrate the relationship between variables in more clearly (Graph 3). As we can see this line have a positive angle that means that on average the greater Height respondent has the greater weight he has too¹⁰.



GRAPH 3. SCATTER-PLOT FOR HEIGHT AND WEIGHT FEATURES

5.3 BUILD A LINEAR REGRESSION OF ONE OF THE FEATURES OVER THE OTHER. MAKE A COMMENT ON THE MEANING OF THE SLOPE.

To build a linear regression we defined Weight feature as dependent variable and Height feature as independent. Regression equation looks in the following way:

¹⁰ All graphs and calculations of this section were made in R software ([https://en.wikipedia.org/wiki/R_\(programming_language\)](https://en.wikipedia.org/wiki/R_(programming_language)))

$$Weight = \alpha_0 + \alpha_1 * Height + \varepsilon_i \quad (1)$$

In the next step, estimates of the regression model parameters were obtained. According to the obtained results, the regression equation is as follows :

$$Weight = -119,34 + 1,069 * Height \quad (2)$$

Interpretation of the coefficient α_1 - at a higher value of the growth variable by one unit on average the weight variable becomes 1,069 more. Geometrically, α_1 is the tangent of the slope of direct regression and as we can see in equation (2) $\alpha_1 = 1,069$. It means that the slope of the regression line is positive and has a little bit more than 45° (about 46-47) and Graph 3 above illustrates the same result.

5.4 FIND THE CORRELATION AND DETERMINACY COEFFICIENTS, AND COMMENT ON THE MEANING OF THE LATTER.

Correlation coefficient can be estimated as follow:

$$r_{xy} = \frac{\overline{x*y} - \bar{x} * \bar{y}}{S(x) * S(y)} \quad (3)$$

where r_{xy} - correlation coefficient between features x and y, $\overline{y * x}$, \bar{y} , \bar{x} - the averages , and $S(x)$, $S(y)$ - standard deviations of features. In our case correlation coefficient can be presented as follow (h for Height and w for Weight):

$$r_{hw} = \frac{\overline{h*w} - \bar{h} * \bar{w}}{S(h) * S(w)} = 0,7297029 \quad (4)$$

Obtained meaning of the correlation coefficient shows that the features we analyzed have strong positive (or direct) correlation in terms of Cheddok scale.

~~Let's move on further for the calculation of determination coefficient. The determination coefficient shows the proportion of the variance of real y explained by the linear regression. Its square root, ρ , is referred to as the coefficient of~~

multiple correlation between real y we observe (Weight in our case) and $X = \{x_0, x_1, x_2, \dots, x_p\}$ (Height in our case).

In our example, determination coefficient $p^2 = 0,5325$, that is, the Height feature explain about 53% of the variance of Weight feature, and the multiple correlation is $\rho = 0,5325$ too.

5.5 MAKE A PREDICTION OF THE TARGET VALUES FOR GIVEN TWO OR THREE PREDICTOR' VALUES; MAKE A COMMENT

We took three observations with following values of dependent feature and the corresponding independent variable values:

$$\begin{aligned} Weight_i &= (48, 57, 67), i \in \{1, 2, 3\} \\ Height_i &= (163, 163, 176), i \in \{1, 2, 3\} \end{aligned}$$

Estimated values with linear regression parameters were obtained this way:

$$Weight_1 = -119,34 + 1,069 * Height_1 = -119,34 + 1,069 * 163 = 54,907$$

$$Weight_2 = -119,34 + 1,069 * Height_2 = -119,34 + 1,069 * 163 = 54,907$$

$$Weight_3 = -119,34 + 1,069 * Height_3 = -119,34 + 1,069 * 176 = 68,804$$

Below (Table 11) you can see deviations of estimated values from real values in units and percentages:

# of observation	Real value	Estimated value	Deviation (abs.)	Deviation (abs.), %
1	48	54,9	6,9	14,4%
2	57	54,9	2,1	3,7%
3	67	68,8	1,8	2,7%

For two observations with different real dependent variable values we got the same meaning of estimated y as the values of independent variable were the same so we can conclude that linear regression implies high level of approximation. In average errors are not vary significant so that we observe about 2-3% deviation in two cases and we can say that in our example of such features like height and weight linear regression is pretty good tool to understand the relation.

5.6 COMPARE THE MEAN RELATIVE ABSOLUTE ERROR OF THE REGRESSION ON ALL POINTS OF YOUR SET AND THE DETERMINACY COEFFICIENT AND MAKE COMMENTS

As we mentioned above determinants coefficient for our regression is equal to 0,5325. The relative absolute error for a given observations (MRAE) = 1,255



Applications

Code in MatLab for PCA

%Given X and its centered version Y

X= fearpca;

[nn,mm]=size(X);

[me,range,mmin,mmax,sst] = stand(X);

% Y=(X-repmat(me,nn,1))./repmat(range,nn,1); %standardised data

Y=(X-repmat(mean(X),nn,1))./repmat(std(X),nn,1); %standardised data

[Z,Mu,C0]=svd(Y);

Z=Z(:,1:6) % three singular 6D scoring vectors

mu=diag(Mu)

ds=sum(sum(Y.*Y)) % data scatter

% z = -z; c = -c;

contr1=mu(1)^2

contr2=mu(2)^2

contr3=mu(3)^2

contr4=mu(4)^2

contr5=mu(5)^2

contr6=mu(6)^2

contr1percent=mu(1)^2/ds

contr2percent=mu(2)^2/ds

contr3percent=mu(3)^2/ds

contr4percent=mu(4)^2/ds

contr5percent=mu(5)^2/ds

contr6percent=mu(6)^2/ds

% Compute the covariance matrix B, first eigenvalue and eigenvector of B

B=Y'*Y/nn;

[C, La]=eig(B);% eigenvalues in the descending order

La=diag(La)

c1=C(:,6)

c2=C(:,5)

c3=C(:,4)

c4=C(:,3)

c5=C(:,2)

c6=C(:,1)

% Given centered data Y, eigenvalue La1 and eigenvector c1, compute the Principal component scoring vector

z1=Y*c1/sqrt(677*La(6));

z2=Y*c2/sqrt(677*La(5));

z3=Y*c3/sqrt(677*La(4));

z4=Y*c4/sqrt(677*La(3));

z5=Y*c5/sqrt(677*La(2));

z6=Y*c6/sqrt(677*La(1));

plot(Z(:,1),Z(:,2),'k.')

% Hidden factor score

ma=max(X);

Y1=X*100./repmat(ma, 677, 1)

[z0,mu0,c0]=svd(Y1);

c1=-c0(:,1)

alpha=1./sum(c0);

100*mu0(1,1)^2/sum(sum(Y1.*Y1))