Исследование ППС Уральского федерального университета

Смирнов В.А.

Раздел I. Изучение данных

Прежде чем анализировать данные стоит понять как они устроены, имеются ли пропуски и т.д.

Таблица1 Структура данных

```
dd=data_frame(name=dd$name, dol=dd$dol, step=dd$step)

dd %>%
  rename ('ФИО'=name, 'Должность'=dol, 'Степень'=step) %>%
  datatable()
```

Show 10 **entries** **Search:**

	ФИО	Должность	Степень
1	Абаимов Николай Анатольевич	Инженер 1 категории, Ассистент	
2	Абдуллин Ренат Рашидович	Ведущий инженер, Старший преподаватель	кандидат технических наук
3	Аболина Татьяна Михайловна	Доцент	кандидат филологических наук
4	Абрамов Александр Валерьевич	Доцент, Старший научный сотрудник	кандидат химических наук
5	Абрамова Софья Борисовна	Доцент	кандидат социологических наук
6	Абржина Лариса Леонидовна	Доцент	кандидат экономических наук
7	Аввакумова Екатерина Анатольевна	Научный сотрудник, Доцент	кандидат физико-математических наук
8	Авдеева Вера Владимировна	Заместитель директора, Доцент	Кандидат искусствоведения

	ФИО	Должность	Степень
9	Аверкова Ольга Владимировна	Старший преподаватель	
10	Автохутдинова Ольга Федоровна	Доцент	кандидат филологических наук

Showing 1 to 10 of 1,828 entries

Previous 1 2 3 4 5 … 183 Next

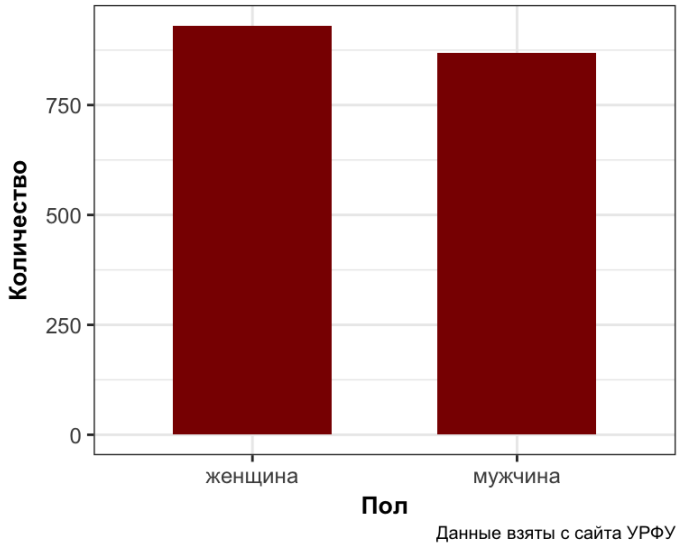
```
#визуализация структуры данных
g1=vis_dat(dd)
g2=vis_miss(dd)

#plot_grid(g1, g2, ncol = 1,  labels = c ('Type', 'NA'))
```

```
# уберем повторы
dd=dd %>%
  distinct()
```

Раздел II. Начало анализа: социально-демографический анализ

```
# добавим столбец с полом
dd$oo=ifelse (str_sub(dd$name, -1)=="а", "женщина", "мужчина")


ggplot (dd, aes(oo))+
  geom_bar(width = 0.6, fill='darkred')+theme_bw()+xlab ('Пол')+ylab('Количество')+
  labs (caption='Данные взяты с сайта УРФУ')+ggtitle('Рис.1 Распределение ППС УРФУ по полу')+
  #geom_text(x=2, y=77, label=paste (ch1$statistic,round (ch1$p.value, 2), sep = '\n'),
            #check_overlap = T, size=2.5)+
   theme(axis.text = element_text(size=9),
         axis.title = element_text(size=10, face='bold'),
         plot.caption = element_text(size=7.5),
         plot.title = element_text(size=11, face = 'bold', hjust = 0.5))
```

Рис.1 Распределение ППС УРФУ по полу




```
#gganimate(g1)

detach ('package:desctable')

ch1=chisq.test(table (dd$oo))
```

Тест Хи-квадрат позволяет отвергнуть гипотезу о значимости различий между мужчинами и женщинами. Значение критерия равняется - 2.2062257, p-value - 0.1374546

Раздел III. Продолжение анализа: изучение преподавательских сетей

Поищем семейные связи в академической среде.

Рис.2 Кластеры однофамильцев

```
#разделим фио на три столбца
dd1=dd %>%
  separate(name, into=c ('fam', 'nam', 'ot'))


# создадим датафрайм с общим столбцом фамиий
female=dd1 %>%
  filter (oo=='женщина') %>%
  mutate (ff=ifelse(str_sub(fam, -1)=='а', str_sub(fam, 0, -2), str_sub(fam, 0, -1)))

male=dd1 %>%
  filter (oo=='мужчина') %>%
  mutate(ff=fam)

dd2=rbind(female, male)

#отберем наиболее часто встрающиеся фамилии
d1=dd2 %>%
  count (ff, sort=T) %>%
  filter (n>7) %>%
  select(ff)

f=str_c(d1$ff, collapse = '|')

f='Попов|Иванов\\b|Ермаков|Кузнецов|Степанов\\b|Иванова\\b|Степанова\\b'

dd3=dd2 %>%
  filter (str_detect(fam, f)) %>%
  arrange(ff) %>%
  select(ff, nam, ot) %>%
  unite (c(nam, ot), sep=' ', -ff) %>%
  graph_from_data_frame()

#построим график
ggraph(dd3, layout = 'fr')+
  geom_edge_link(linetype=3) +
  geom_node_point(color = "blue", size = 2, alpha=0.8) +
  geom_node_label (aes(label = name),  size=2.3, repel = T) +
  theme_bw()+
  theme (axis.ticks = element_blank(),
         axis.text = element_blank(),
         axis.title = element_blank())
```

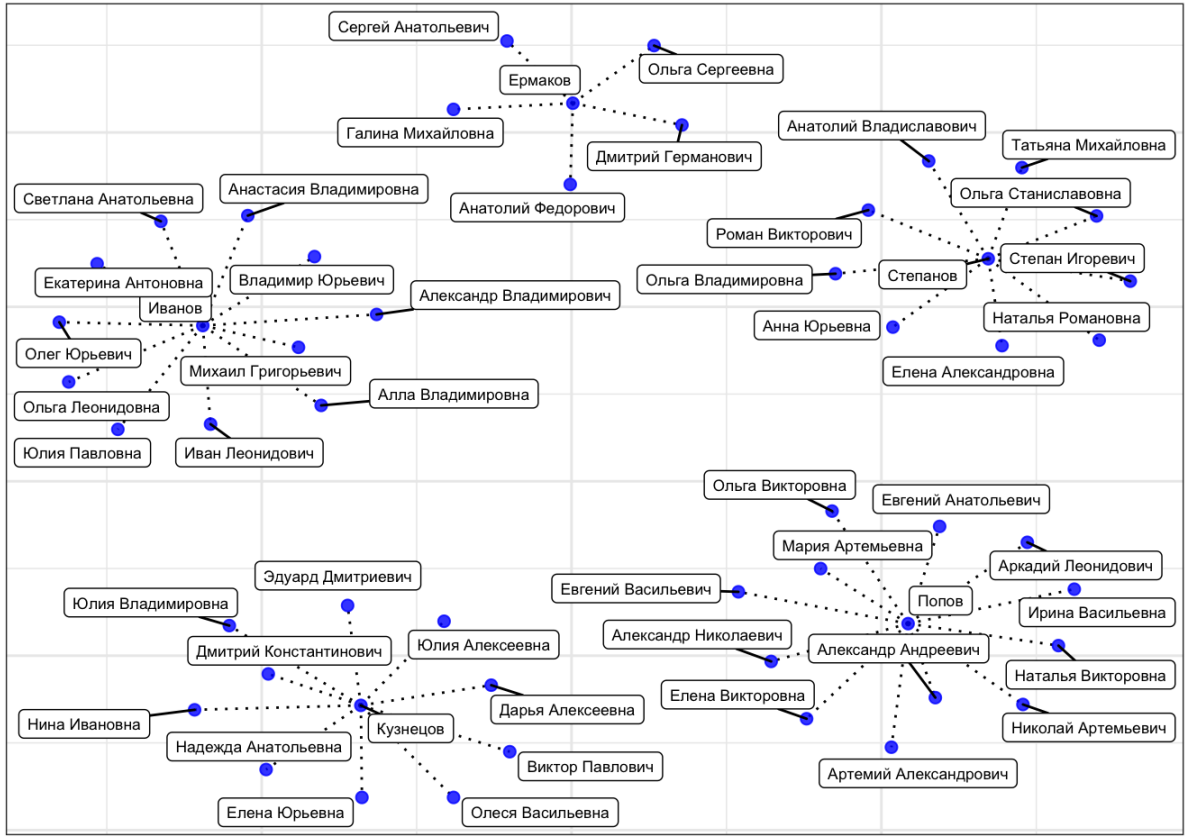


Таблица 2 Наиболее часто встречающиеся фамилии ППС

```
#представим в таблице, наиоболее часто встреающиеся фамилии
#отберем наиболее часто встрающиеся фамилии
library(desctable)

d1=dd2 %>%
  count (ff, sort=T) %>%
  filter (n>5) %>%
  select(ff)

f=str_c(d1$ff, collapse = '|')

dd2 %>%
  select(fam, nam, ot, dol, step, ff) %>%
  filter (str_detect(fam, f)) %>%
  arrange(ff) %>%
  unite (col = 'fio', fam, nam, ot, sep=' ') %>%
  select(ff, fio, dol, step) %>%
  rename ('Кластер'=ff, 'ФИО'=fio, 'Должность'=dol, 'Степень'=step) %>%
  datatable()
```

Show 10 entries Search:

	Кластер	ФИО	Должность	Степень
1	Волков	Волкова Анна Альбертовна	Доцент	кандидат технических наук
2	Волков	Волкова Марина Владимировна	Ведущий специалист по аналитической работе, Старший преподаватель	
3	Волков	Волкова Надежда Евгеньевна		кандидат химических наук
4	Волков	Волкова Яна Юрьевна	Доцент, Специалист по сопровождению образовательных программ	Кандидат физико-математических наук
5	Волков	Волков Аркадий Германович	Доцент, Ведущий научный сотрудник	к.ф.-м.н
6	Волков	Волков Владимир Алексеевич	Доцент	кандидат физико-математических наук

	Кластер	ФИО	Должность	Степень
7	Волков	Волков Михаил Владимирович	Главный научный сотрудник, Заведующий кафедрой	д.ф.м.н.
8	Волкович	Волкович Владимир Анатольевич	Доцент	PhD (1998), кандидат химических наук (2004)
9	Иванов	Иванова Алла Владимировна	Доцент, Ведущий научный сотрудник	кандидат химических наук
10	Иванов	Иванова Анастасия Владимировна	Доцент, Директор	кандидат экономических наук

Showing 1 to 10 of 64 entries

Previous 1 2 3 4 5 6 7 Next

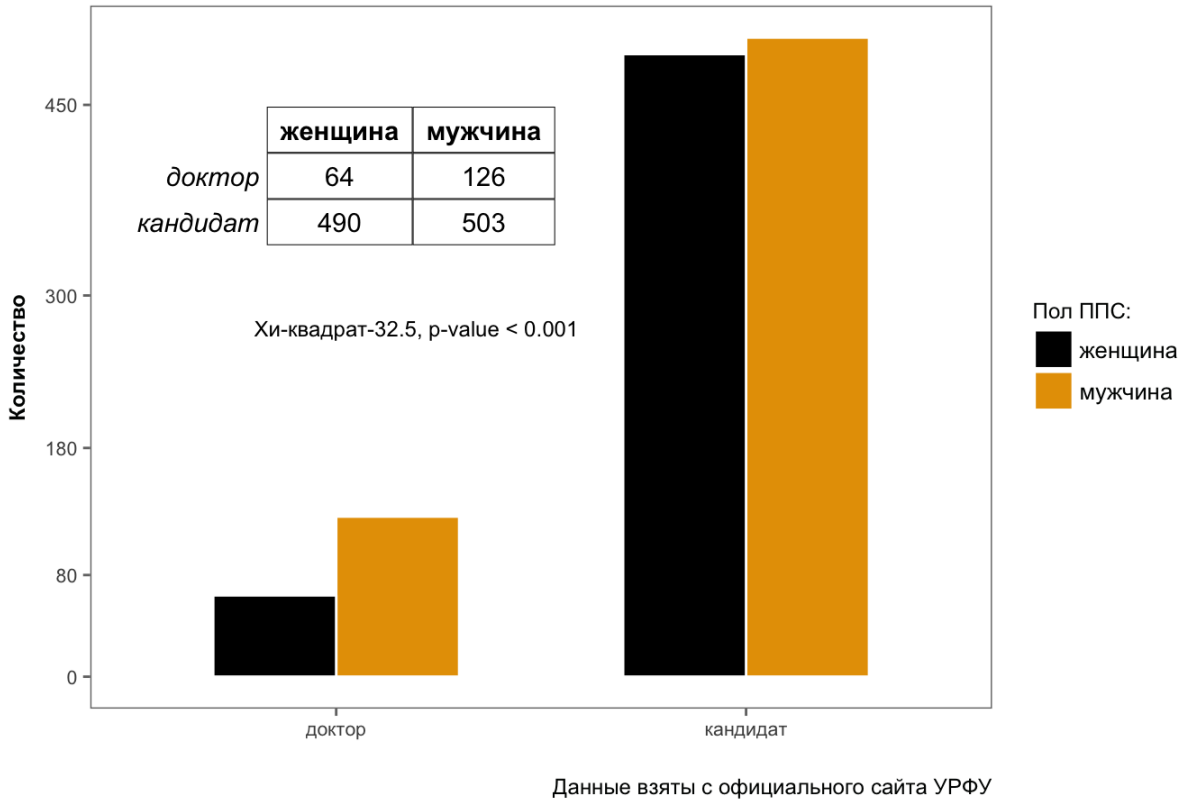
Раздел IV. Научная квалификация ППС УРФУ

Проанализируем распределение ученых степеней ППС в зависимости от пола.

```
dd4=dd %>%
  separate(dol, into=c ('d1', 'd2'), sep=',')


#dd4 %>%
  #select(d1, d2, step, oo) %>%
  #count (d1, sort = T)


#dd4 %>%
  #select (d1, step, name) %>%
  #filter(d1=='Доцент' & str_detect(step, '[Дд]октор')==T)


#dd4$step=tolower(dd4$step)

dd5=dd4 %>%
  separate(step, into=c ('s1', 's2', 's3'), na.rm=T) %>%
  select(s1, s2, s3, oo)


#dd5$s2[is.na (dd5$s2)] ='Без степени'




dd5=na.omit(dd5)


dd5$ss1=ifelse(str_sub (dd5$s1, 1,1)=='д', 'доктор', 'кандидат')


#dd5$ss1=ifelse(str_sub (dd5$s1, 1,1)=='д', 'доктор', ifelse(str_sub (dd5$s1, 1,1)=='б',  'без степени', 'кандидат'))

#chisq.test(t)

t=table (dd5$ss1, dd5$oo)

t1=ggtexttable(t, theme = ttheme ('classic'))

p1=ggparagraph('Хи-квадрат-32.5, p-value < 0.001', size = 9)
```

```
g1=ggplot (dd5, aes(ss1, fill=oo))+
  geom_bar(width = 0.6, position = 'dodge', col='white')+theme_few()+
  labs (x='', y='Количество', caption='Данные взяты с официального сайта УРФУ')+
  ggtitle('Рис.3 Распределение научных степеней по полу')+
  scale_fill_colorblind(name='Пол ППС:')+
  scale_y_continuous(breaks = c (0, 80, 180, 300, 450))+
  theme(axis.title = element_text(size=9, face='bold'),
        axis.text = element_text(size=8),
        plot.caption = element_text(size=9),
        plot.title = element_text(size=11, face='bold', hjust = 0.5),
        legend.title = element_text(size=9))


g2=g1+annotation_custom(ggplotGrob(t1),
                                xmin = 1, ymin = 260,
                                xmax = 1)


g2+annotation_custom(ggplotGrob(p1),
                     xmin=0.8, xmax = 1.6 ,
                     ymin=185, ymax = 280)
```

Рис.3 Распределение научных степеней по полу



Раздел V. Исследование социальных сетей

Используя API vkontakte найдем сообщества, имеющие отношение к УРФУ. Работать с социальной сетью можно, используя специализированные библиотеки или же непосредственно создавая запросы, которые передаются на сервер. Мы будем использовать второй вариант.

Таблица 3 Сообщества, связанные с УРФУ

```
s='https://api.vk.com/method/groups.search?q=УРФУ&fields=description&v=5.52&access_token=ef1e39b50855209d819379289dcd9ab255e4620e222b55be3a30be8f5a141f0e91a54b3098a46d3a68816'


res=GET (s)

# преобразум json в таблицу
rr=httr::content(res, as='text')
rr=fromJSON(rr)

rr=as.data.frame(rr)

rr %>%
  filter (str_detect(response.items.name, 'ФУ')) %>%
  select(name=response.items.name, opis=response.items.description, id=response.items.id) %>%
  datatable()
```

Show 10 entries

Search:

	name	opis	id
1	Уральский федеральный университет | УрФУ	Официально, Уральский федеральный. Привет! Еще мы есть здесь: 💡 t.me/urfu_ru 💡 facebook.com/ural.federal.university 💡 twitter.com/urfu 💡 instagram.com/urfu.ru 💡 ok.ru/uralfederal Анонсы событий #УрФУ: http://urfu.ru/ru/events/ Правила: 1. За мат, спам и оскорбления в комментариях — бан. 2. Мы не размещаем рекламу. Все вопросы пишите на partner@urfu.ru 3. Мы не размещаем информацию, не связанную с университетом.	22941070
2	СТУДЕНТ УрФУ	Специально для всех, кто учится и работает в УрФУ, студенческая редакция каждый день создаёт атмосферу, а в перерывах рассказывает новости университета и разные полезности.	21604404
3	========== УрФУ ==========	Вот и наступил новый этап в истории нашего университета... теперь мы УрФУ!	45925

	name	opis	id
4	Хочу в УрФУ | Абитуриент УрФУ	Вся важная информация о поступлении в Уральский федеральный теперь в одном месте! Информация о приеме: https://urfu.ru/ru/applicant/ Направления подготовки бакалавриата и специалитета: https://urfu.ru/ru/applicant/docs-abiturient/programs/ Направления подготовки магистратуры: https://magister.urfu.ru/ru/programs/ Направления подготовки аспирантуры: https://aspirant.urfu.ru/ru/aspirantura/dlja-postupajushchikh/napravlenijaprofili-podgotovki/ Телефон контакт-центра УрФУ: 8-800-100-50-44 (звонок по России бесплатный) +7(343)375-44-44 Приемная комиссия: +7 (343) 375-44-74, priem@urfu.ru г. Екатеринбург, ул. Мира, 19, каб. ГУК-100 Принимаем звонки и сообщения по будням с 8:30 до 19:00 и в субботу с 10:00 до 16:00 Волнующие вопросы задавайте в сообщения сообщества.	22301031
5	Подслушано в УрФУ (Екатеринбург)		59432260
6	POS News УрФУ | Будь в курсе	Делаем круто вместе!	24827886
7	Спорт УрФУ	Рассказываем о спортивных событиях Уральского федерального.🏆 Подробная информация о грядущих событиях, новости спорта, интервью с известными спортсменами Уральского федерального, а также репортажи с самых интересных мероприятий теперь на официальной странице Центра спортивно-массовой и оздоровительной работы УрФУ. Наш хэштег #яумамычемпион	76308265
8	Магистрант УрФУ	Официальная группа для магистрантов Уральского федерального университета им. Б.Н. Ельцина и всех тех, кто хочет стать ими! Магистратура Уральского федерального университета дает возможность выпускникам бакалавриата и специалитета других вузов увеличить число компетенций в рамках выбранных направлений подготовки и углубить свое базовое образование. Это прекрасный шанс получить второе высшее образование на бюджетной форме обучения, по направлению отличному от того, на котором Вы получили или получаете диплом бакалавра или специалиста. Два разноплановых диплома о высшем образовании за 6 лет обучения оценит любой работодатель! Многочисленные программы международного обмена и зарубежные стажировки, мощнейшая лабораторная база и связи с ведущими предприятиями Урало-Сибирского региона станут прочнейшей основой Вашего будущего! FB https://www.facebook.com/master.urfu	38695732

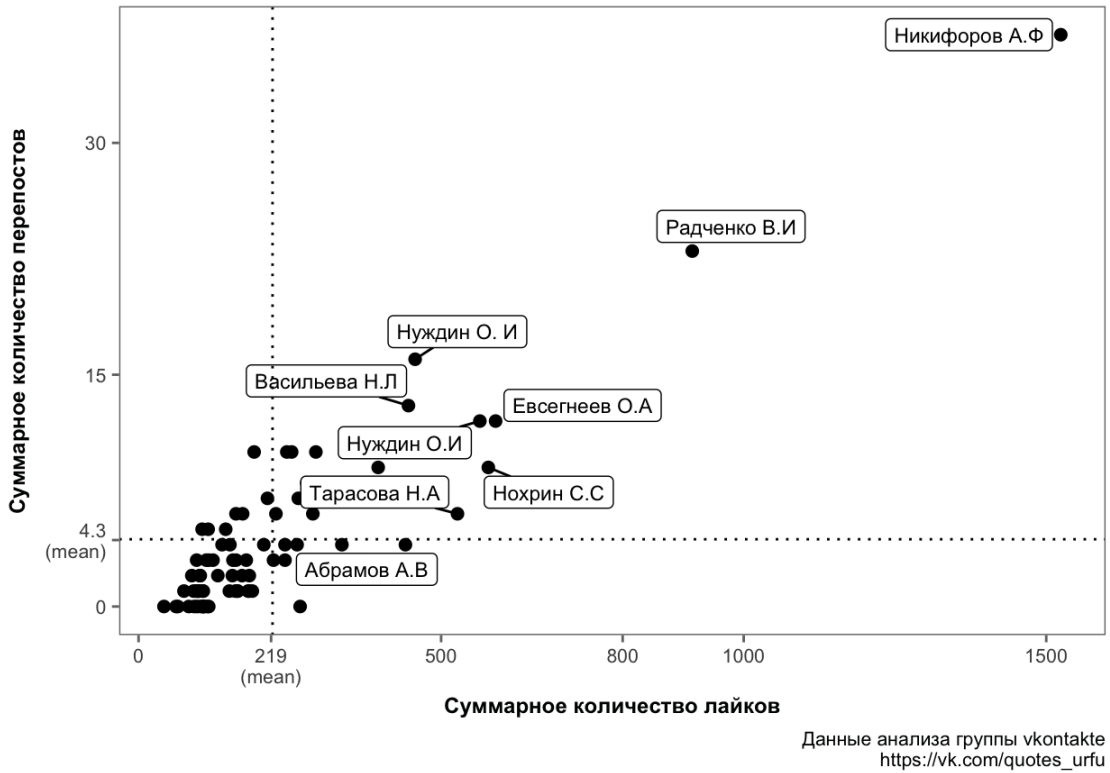
	name	opis	id
9	Цитаты преподавателей УрФУ	Уважаемые подписчики! Добро пожаловать! Здесь - только самые интересные, самые смешные, и самые мудрые высказывания наших любимых преподавателей! Предлагайте цитаты через функцию "Предложить новость". Изречения будут публиковаться регулярно. И сразу небольшая просьба. Пожалуйста, указывайте вместе с цитатой фамилию автора и соответствующий институт УрФУ. Те , кто сразу вставляет в новость хештеги (фамилия автора и институт), получают жирный плюсик к карме! Спасибо, что уменьшаете объем работы для администрации! Пример оформления цитаты: "Самореализация без промежуточных посредников в нашей стране возможна только в двух случаях: если ты гений, либо везунчик." ИСПН, Петров А.В #ИСПН #Петров В обсуждениях паблика есть список хэштегов: http://vk.com/topic-77294597_30537054 Спасибо! ПОДПИСЫВАЙТЕСЬ И РАССКАЗЫВАЙТЕ ДРУЗЬЯМ!	77294597
10	Фотоклуб УрФУ	Фотоклуб УрФУ — открытое сообщество фотографов Уральского федерального университета. Мы объединяем профессионалов и любителей фотоискусства в одну команду и освещаем мероприятия, проходящие в УрФУ и Екатеринбурге. Наши фоторепортажи востребованы в корпоративной прессе, социальных сетях, а также используются при выпуске литературы и фирменной продукции. В 2014 году мы создали образовательную площадку — открытую [club79193097|Фотошколу] для студентов и сотрудников университета. Самые активные слушатели приглашаются в Фотоклуб для освещения крупных событий в Уральском федеральном. По вопросам фотосъемки вашего события пишите [id11035262|руководителю Фотоклуба Илье Сафарову] Фотоклуб, каким его видите вы: instagram.com/fotocluburfu #ФотоУрФУ	44119943

```
res=GET ('https://api.vk.com/method/wall.get?owner_id=-77294597&count=100&v=5.52&access_token=ef1e39b50855209d819
379289dcd9ab255e4620e222b55be3a30be8f5a141f0e91a54b3098a46d3a68816')


t=httr::content (res, as='text')
t=fromJSON(t, flatten = T)

t=as.data.frame(t)

rr=httr::content(res, as='text')

rr=fromJSON(rr)

l=rr$response$items$likes %>%
  select(count) %>%
  rename (l_count=count)

r=rr$response$items$reposts %>%
  select(count) %>%
  rename (r_count=count)

d=rr$response$items$date
txt=rr$response$items$text


data=cbind(txt, d, l, r)



#data %>%
  #arrange(desc(r_count))

#извлекаем фамилии преподавателей

f1='[А-Я][а-я]+\\s+[А-Я]\\.+\\s{0,2}[А-Я]' # фамилия стоит первой
f2='\\s+[А-Я]\\.+\\s{0,2}[А-Я]\\.\\s{0,2}[А-Я][а-я]+\\s' # инициалы стоят первыми

#приведем все к одному виду
g=unlist (str_extract_all(data$txt, f2))
g=str_replace_all(g, ' ', '')

df=data_frame(g1=str_sub(g, 5, -1),
              g2=str_sub(g, 1,4))
```

```
df=df %>%
  unite(g, g1,g2, sep=' ')
gg=as.vector(df$g)




#заменим фио
data$txt=str_replace_all(data$txt, f2, gg)

data=data %>%
  filter (str_detect(txt, f1)) %>%
  mutate(fio=unlist (str_extract_all(txt, f1)))

data1=data %>%
  group_by(fio) %>%
  summarise(sum_l=sum (l_count),
            sum_r=sum (r_count))
```

```
ggplot (data1, aes(x=sum_l, y=sum_r))+
  geom_point(col='black', size=2)+
  geom_label_repel(aes(label=fio), data = filter (data1, sum_l>400), size=3, inherit.aes = T)+theme_few()+
  xlab ('Суммарное количество лайков')+ylab ('Суммарное количество перепостов')+
  geom_vline(xintercept = mean (data1$sum_l), linetype=3)+
  geom_hline(yintercept = mean (data1$sum_r), linetype=3)+
  scale_x_continuous(breaks = c (0, 219, 500, 800, 1000, 1500),
                     labels = c ('0', '219\n(mean)', '500', '800', '1000', '1500'))+
  scale_y_continuous(breaks = c (0, 4.3, 15, 30),
                     labels = c ('0', '4.3\n(mean)', '15', '30'))+
  labs (caption='Данные анализа группы vkontakte\nhttps://vk.com/quotes_urfu')+
  ggtitle('Рис.4 Распределение наиболее популярных ППС\nпо количеству лайков и перепостов')+
  theme (axis.title = element_text(size=9, face='bold'),
         axis.text = element_text(size=8),
         plot.caption = element_text(size=8),
         plot.title = element_text(size=11, face='bold', hjust = 0.5))
```

Рис.4 Распределение наиболее популярных ППС по количеству лайков и перепостов



Построим предсказательную модель, описывающую влияние количество лайков на число перепостов. Для этого создадим регрессионную модель.

Таблица 3 Результаты линейной регрессии

```
fit=lm (data1$sum_r~data1$sum_l)
summary(fit)
```

```
##
## Call:
## lm(formula = data1$sum_r ~ data1$sum_l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7717 -1.4623 -0.3144  1.3445  6.3866
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.024164   0.404894  -2.529   0.0136 *
## data1$sum_l  0.024281   0.001314  18.475   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.422 on 72 degrees of freedom
## Multiple R-squared:  0.8258, Adjusted R-squared:  0.8234
## F-statistic: 341.3 on 1 and 72 DF,  p-value: < 2.2e-16
```

Проанализируем встречаемость слов в цитатах ППС, построим облако тегов. Изучим процесс стемминга, токенизации.

```
dd=data %>%
  select (txt)

#извлечем текст, без хэштэгов
#str_view_all(data$txt, '\\w*')

s='\\w*'
ttt=unlist (str_extract_all(data$txt, s))

ttt=as.data.frame(ttt)

tt=ttt %>%
  filter (str_length(ttt)>3)

#write.table(file='wr.csv', tt, row.names = F)
```

```
# морфологический анализ текста
/Applications/mystem -ldn wr.csv wr.txt
```

Рис 5. Облако тегов, построенное из записей группы

```
rr=read.table('wr.txt')
rr=filter (rr, !str_detect(rr$V1, '[\\s\\?]'))

r=rr %>%
  count (V1, sort=T)

r=as.data.table(r)

#r=r[!r$V1 %in% stop$V1]

wordcloud(r$V1, r$n, min.freq = 3, colors=brewer.pal(8, "Dark2"), random.order = F)
```