# Анализ уровня потребления алкоголя школьниками солнечной Португалии

Герман Соколов

8 апреля 2016 года

## 1. Описание набора с данными.

Источник - репозиторий UC Irvine ('http://archive.ics.uci.edu/ml/').

Количество объектов - 1044 (уникальных - 662).

Количество признаков - 32.

Описание признаков:

- school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
- sex - student's sex (binary: 'F' - female or 'M' - male)
- age - student's age (numeric: from 15 to 22)
- address - student's home address type (binary: 'U' - urban or 'R' - rural)
- famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
- Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
- Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
- Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
- Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
- guardian - student's guardian (nominal: 'mother', 'father' or 'other')
- traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
- studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- failures - number of past class failures (numeric: n if 1<=n<3, else 4)
- schoolsup - extra educational support (binary: yes or no)
- famsup - family educational support (binary: yes or no)
- paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
- activities - extra-curricular activities (binary: yes or no)
- nursery - attended nursery school (binary: yes or no)
- higher - wants to take higher education (binary: yes or no)
- internet - Internet access at home (binary: yes or no)
- romantic - with a romantic relationship (binary: yes or no)
- famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- freetime - free time after school (numeric: from 1 - very low to 5 - very high)
- goout - going out with friends (numeric: from 1 - very low to 5 - very high)
- Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- health - current health status (numeric: from 1 - very bad to 5 - very good)

- absences - number of school absences (numeric: from 0 to 93)
- G1 - first period grade (numeric: from 0 to 20)
- G2 - second period grade (numeric: from 0 to 20)
- G3 - final grade (numeric: from 0 to 20, output target)

## 2. Визуализация данных.

Цель данной работы - эффективно визуализировать данные и найти какие-нибудь интересные закономерности в употреблении спиртных напитков школьниками Португалии.

Для этого загружаем несколько отличных и полезных пакетов:

```
library('ggplot2')
library('RColorBrewer')
library('dplyr')
library('hexbin')
library('gridExtra')
library('reshape2')
library('broom')
```

Далее, выбираем путь к данным и загружаем их, наконец!

Надо заметить, что изначально дано 2 подвыборки - в зависимости от учебного курса, который выбрали школьники (математика либо португальский язык). Поэтому сначала создаем 2 отдельных датафрейма:

```
# There are 2 datasets (Math and Portuguese courses):
filepath_por = "/Users/imac/Dropbox/Study/R/dataset/student-por.csv"
filepath_mat = "/Users/imac/Dropbox/Study/R/dataset/student-mat.csv"
df_por <- read.table(file = filepath_por, sep = ';', header = TRUE)
df_mat <- read.table(file = filepath_mat, sep = ';', header = TRUE)
```

Школьников, которые ходят на оба курса много - 382 человека! Значения признаков для них полностью идентичны за исключением оценок по курсу. Таким образом, можно считать их дубликатами и оставить в единственном экземпляре:

```
# Combine them together and remove duplicate rows (using "unique" features):
df <- rbind(df_por, df_mat)
unique_columns = c("school", "sex", "age", "address", "famsize", "Pstatus", "Medu", "Fedu",
                   "Mjob", "Fjob", "reason", "nursery", "internet")
df <- df[!duplicated(df[ , unique_columns]), ]
```

Данные содержат ни много, ни мало - 3 оценки для каждого школьника (начало, середина, конец курса). Пожалуй, все-таки, это много - найдем средние оценки за курс и выкинем 3 ненужных теперь столбца:

```
df <- mutate(df, aver_grade = (G1 + G2 + G3) / 3)
df <- subset(df, select = -c(G1, G2, G3))
```

Можно также увидеть, что средняя оценка по математике примерно на балл ниже, чем по португальскому языку! Наверное, в моем случае было бы наоборот.

```
df_por <- mutate(df_por, aver_grade = (G1 + G2 + G3) / 3)
df_mat <- mutate(df_mat, aver_grade = (G1 + G2 + G3) / 3)
df_por$course <- 'portuguese'
df_mat$course <- 'math'
means_grades <- rbind(df_por, df_mat)
print(mean(df_por$aver_grade))
```

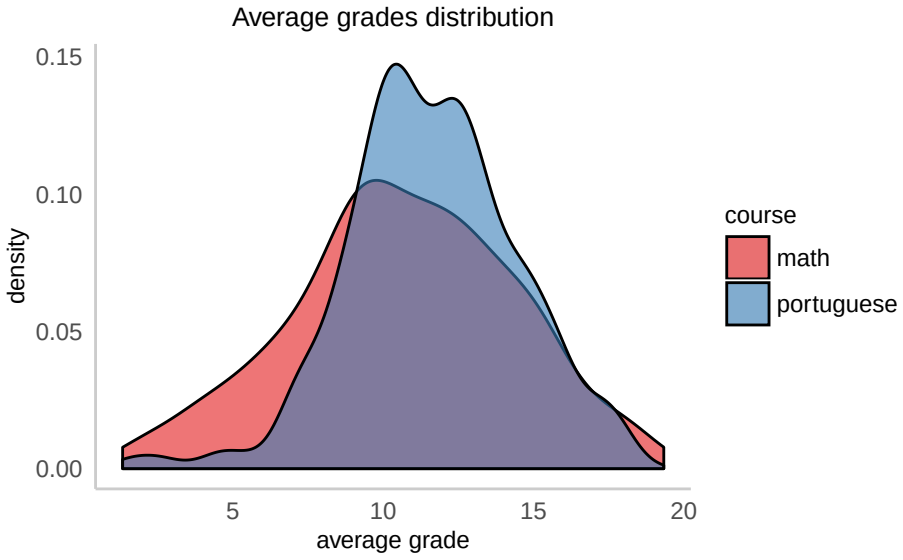
```
## [1] 11.62506
```

```
print(mean(df_mat$aver_grade))
```

```
## [1] 10.67932
```

Построим оценку плотностей функций распределения средних оценок и убедимся, что оценки по математике распределены несколько “левее”:

```
# Compare average grades for different courses:
hist_grades <- ggplot(means_grades, aes(aver_grade, fill = course)) +
               geom_density(alpha = 0.6) +
               ggtitle('Average grades distribution') +
               xlab('average grade') +
               scale_fill_brewer(palette = "Set1", name = "course") +
               theme(panel.background = element_blank(),
                   panel.grid.minor = element_blank(),
                   axis.ticks  = element_blank(),
                   axis.line   = element_line(colour=NA),
                   axis.line.x = element_line(colour="grey80"),
                   axis.line.y = element_line(colour="grey80"),
                   plot.title = element_text(size = 10),
                   axis.title.x = element_text(size = 9),
                   axis.title.y = element_text(size = 9),
                   legend.title = element_text(size = 9))
hist_grades
```

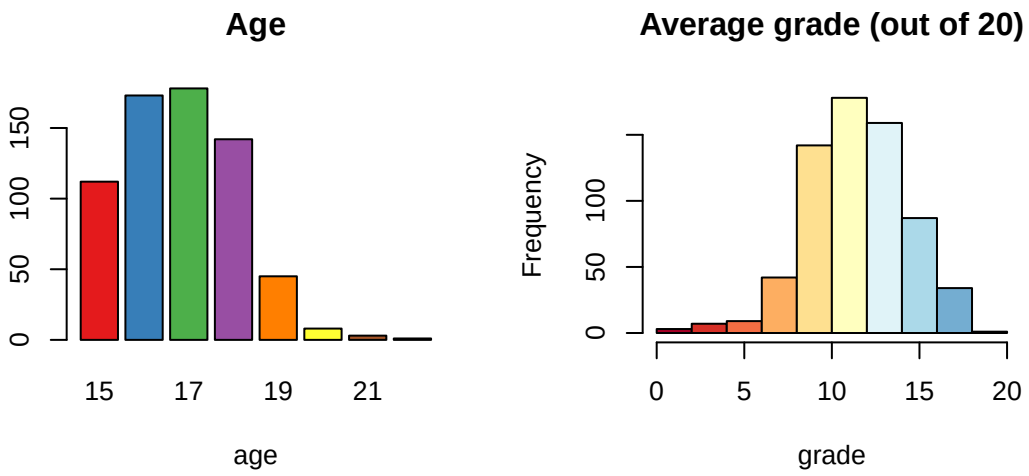


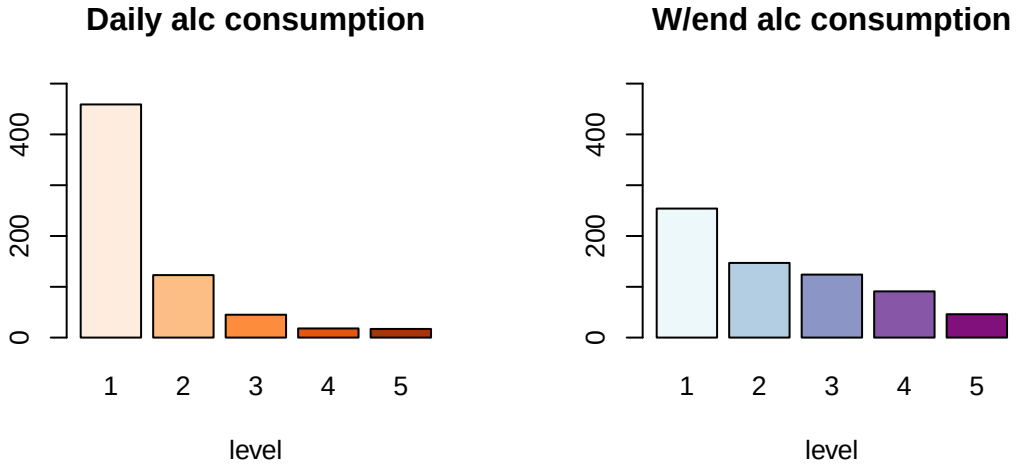
Теперь можно построить несколько гистограмм для объединенного датасета и посмотреть на распределение некоторых признаков.

Например, видно, что:

- Основная часть школьников 15-18 летнего возраста, хотя есть и здоровые “лбы” 22 лет, но это исключение.
- Распределение средних оценок близко к гауссовому.
- Значительная часть школьников не употребляет спиртное в будние дни, но с приходом выходных все меняется и они начинают потихоньку “поддавать”.

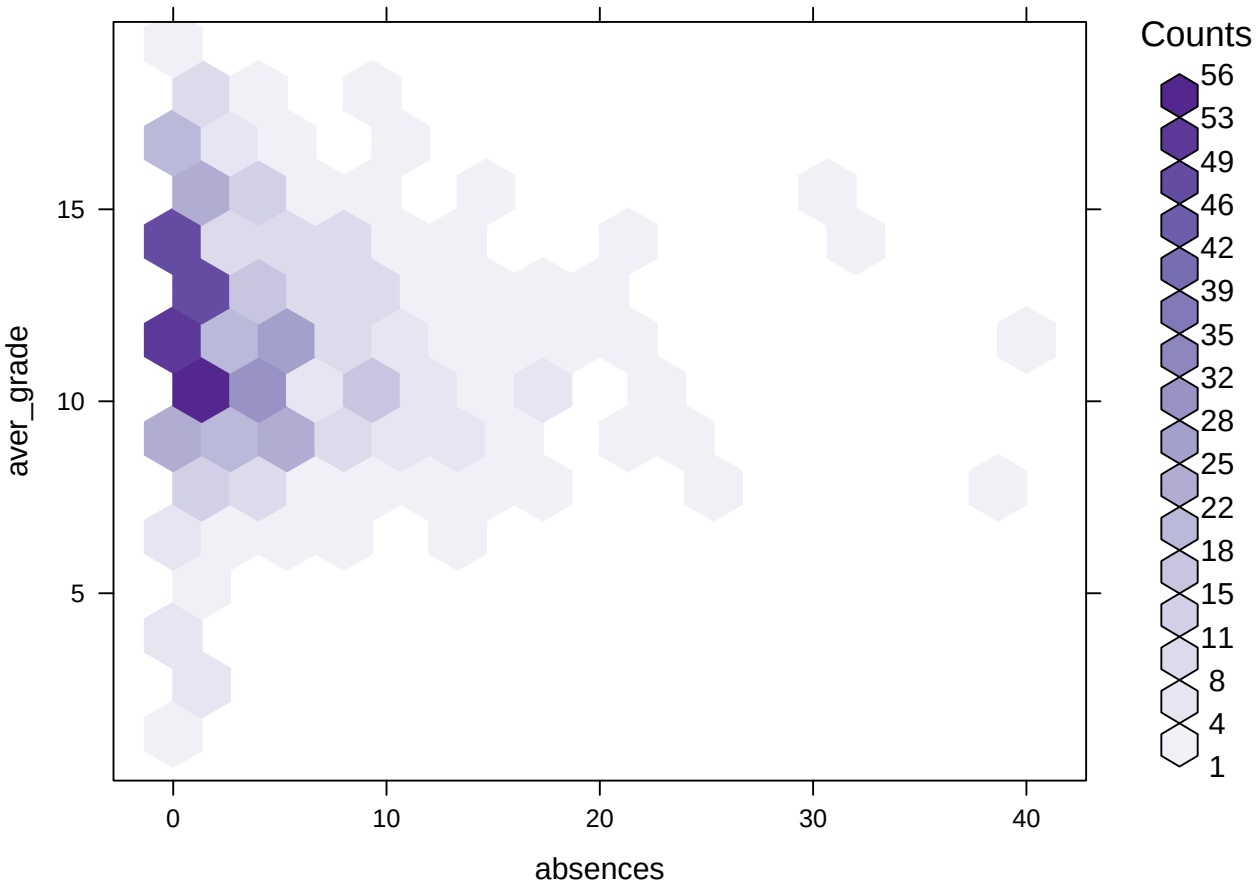
```
# Histograms:
par(mfrow=c(2,2))
barplot(table(df$age), col=brewer.pal(8,"Set1"),
        main="Age", xlab = 'age')
hist(df$aver_grade, breaks=9, col=brewer.pal(11, "RdYlBu"),
     main = "Average grade (out of 20)", xlab = 'grade')
barplot(table(df$Dalc), col=brewer.pal(5, "Oranges"),
        main="Daily alc consumption", xlab = 'level', ylim = c(0, 500))
barplot(table(df$Walc), col=brewer.pal(5, "BuPu"),
        main="W/end alc consumption", xlab = 'level', ylim = c(0, 500))
```



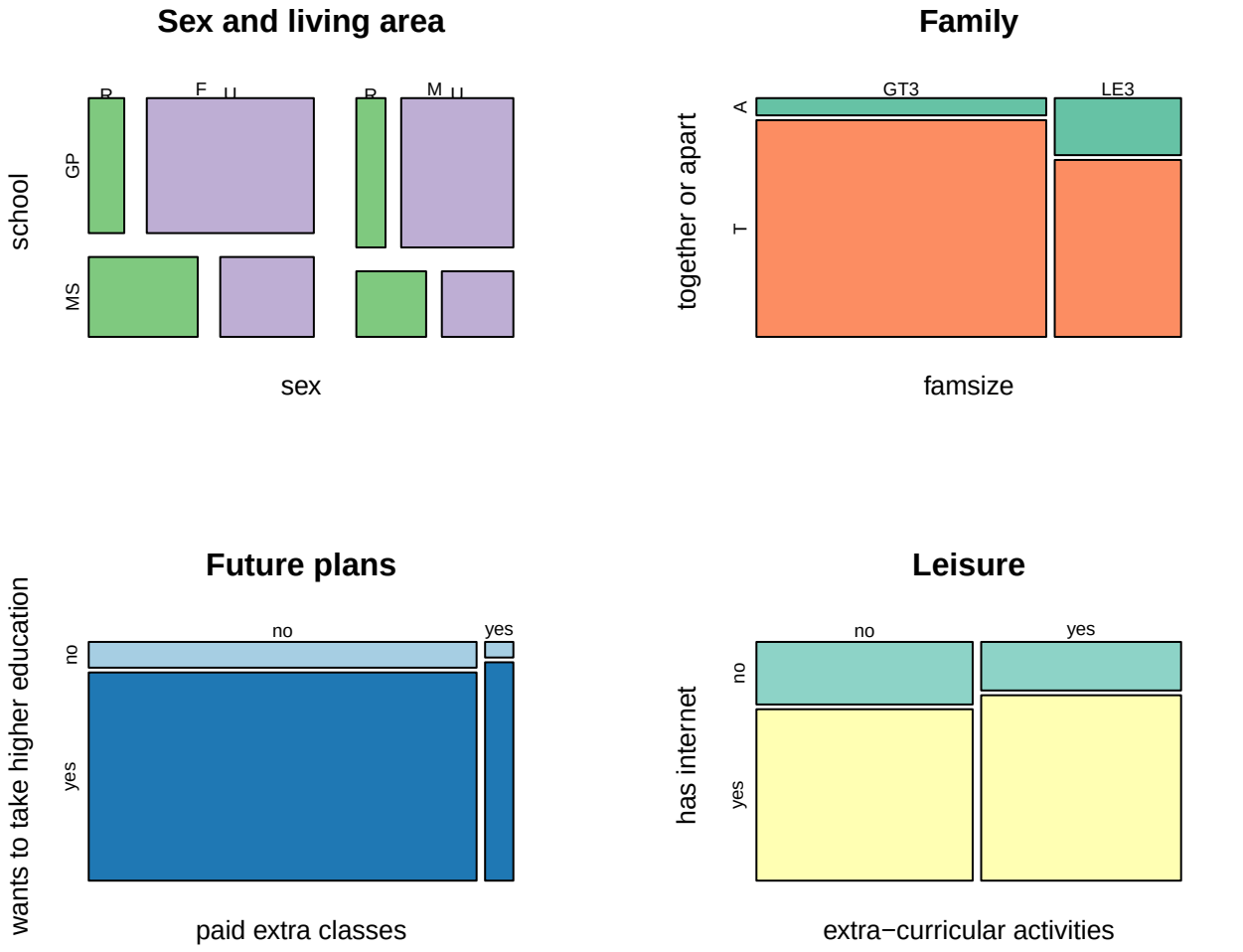
Как пропуски занятий влияют на средние оценки школьников? По графику зависимость непонятна, и если она есть, то незначительная. Подавляющее большинство учеников почти не имеет пропусков, при этом оценки они могут получать самые разные.

```
# Hexbinplot:
rf <- colorRampPalette(brewer.pal(5,'Purples'))
hexb <- hexbinplot(aver_grade ~ absences, data = df, xbins=15, colramp = rf)
hexb
```



Отлично! Настало время “мозаик-плотов” для визуализации нечисловых признаков.

```
# Mosaic plots:
par(mfrow=c(2,2))
mosaicplot(table(df[, c('sex', 'school', 'address')]), color = brewer.pal(3, "Accent"),
         main = 'Sex and living area')
mosaicplot(table(df[, c('famsize', 'Pstatus')]), color = brewer.pal(3, "Set2"),
         main = "Family", ylab = 'together or apart')
mosaicplot(table(df[, c('paid', 'higher')]), color = brewer.pal(4, "Paired"),
         main = "Future plans", xlab = 'paid extra classes', ylab = 'wants to take higher education')
mosaicplot(table(df[, c('activities', 'internet')]), color = brewer.pal(4, "Set3"),
         main = "Leisure", xlab = 'extra-curricular activities', ylab = 'has internet')
```



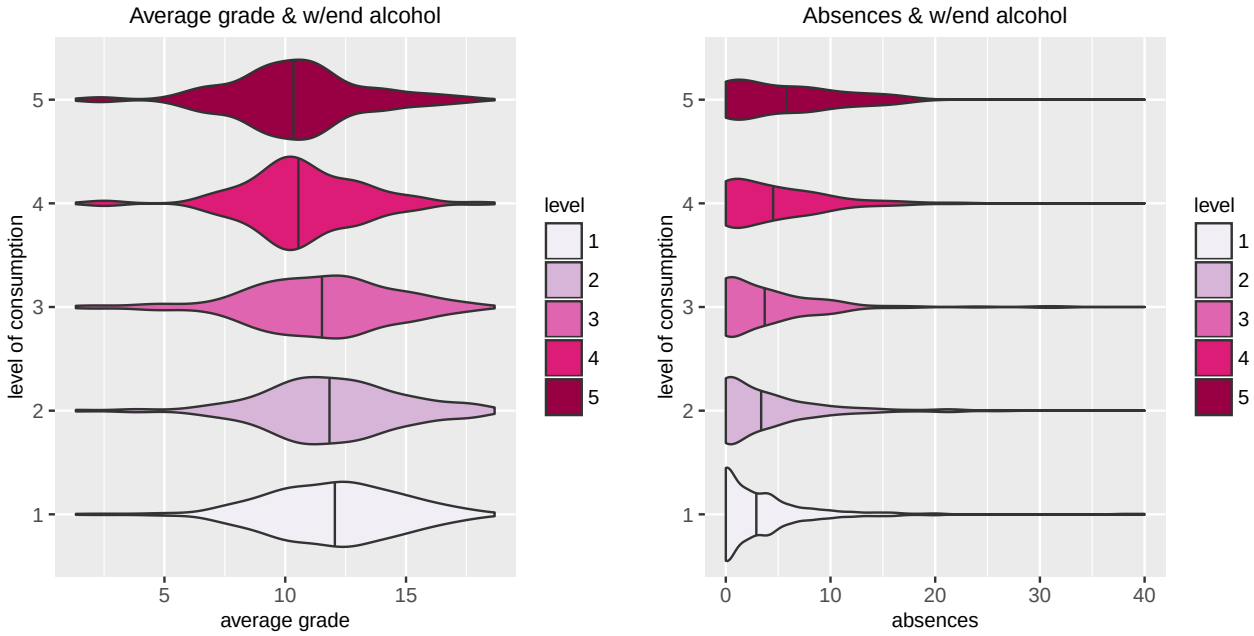

Они дают понимание того, что:

- Школы неоднородны по количеству и составу учеников - в Gabriel Pereira учится около 2/3 всех школьников и подавляющая их чать - дети города, в отличие от другого учебного заведения.
- Около 70% школьников живут в многодетных семьях ($>$ 3 человек) - возможно, поэтому они злоупотребляют. При этом в многодетных семьях вероятнее то, что супруги живут вместе.
- Абсолютное большинство учащихся хотят поступить в универ, при этом такое же большинство не занимается с репетитором. Наличие репетитора, в целом, не влияет на выбор ученика в отношении своего будущего образования.
- Почти у всех есть интернет (2008 год, как ни как) - его наличие скорее способствует тому, чтобы школьник занимался чем-либо дополнительным, кроме уроков.

Замечательно было бы посмотреть, как пропуски учебы и средние оценки влияют на то, насколько много ребенок потребляет алкоголя.

```
# Violin plots:
vp1 <- ggplot() +
      geom_violin(data = df, aes(x = factor(Walc), y = aver_grade, fill = factor(Walc)),
                  draw_quantiles = 0.5, size = 0.5, trim = FALSE) +
      coord_flip() +
      ylab("average grade") + xlab("level of consumption") + ggtitle("Average grade & w/end alcohol") +
      theme(plot.title = element_text(size = 10), axis.title.x = element_text(size = 9),
            axis.title.y = element_text(size = 9), legend.title = element_text(size = 9)) +
      scale_fill_brewer(palette = "PuRd", name = "level")

vp2 <- ggplot() +
      geom_violin(data = df, aes(x = factor(Walc), y = absences, fill = factor(Walc)),
                  draw_quantiles = 0.5, size = 0.5, trim = FALSE) +
      coord_flip() +
      ylab("absences") + xlab("level of consumption") + ggtitle("Absences & w/end alcohol") +
      theme(plot.title = element_text(size = 10), axis.title.x = element_text(size = 9),
            axis.title.y = element_text(size = 9), legend.title = element_text(size = 9)) +
      scale_fill_brewer(palette = "PuRd", name = "level")
grid.arrange(vp1, vp2, ncol = 2)
```

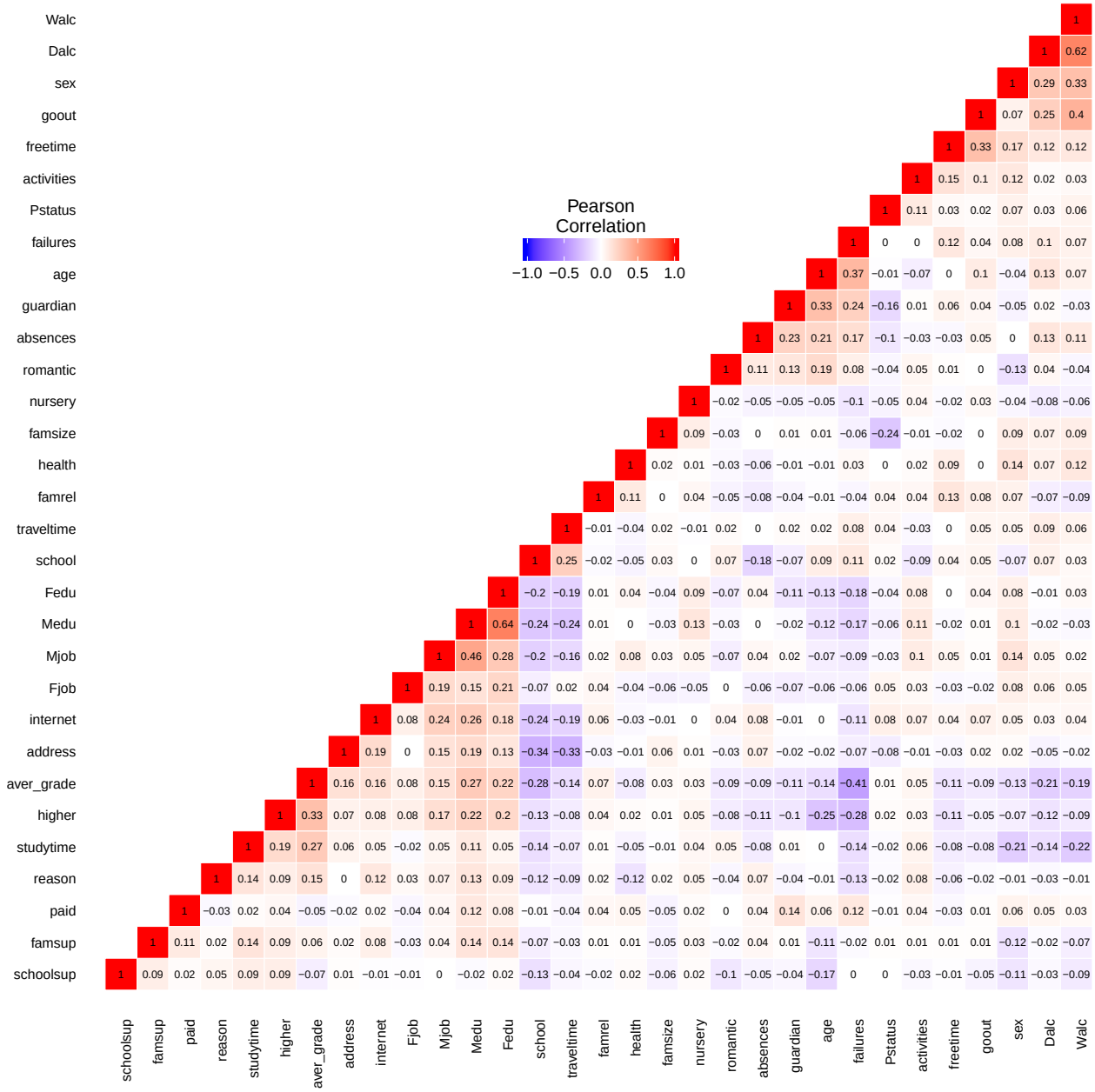



“Виолончелевые-плоты” показывают (в разрезе каждого уровня потребления спиртного) распределение пропусков занятий и оценок (включая медианное значение). Выводы неутешительные:

- Употребеление алкоголя отрицательно связано со средним баллом.
- Чем > пропусков занятий, тем больше выпивает сын/дочка.

На выходе получаем школьника, который не ходит на уроки, получает плохие оценки, и хлыщет спирт по выходным. Конечно, это весьма утрированно, т.к. зависимость пропусков/оценок хоть и присутствует, но незначительная.

Посмотрим на занимательную корреляционную матрицу (код здесь и далее находится в приложении для удобства):



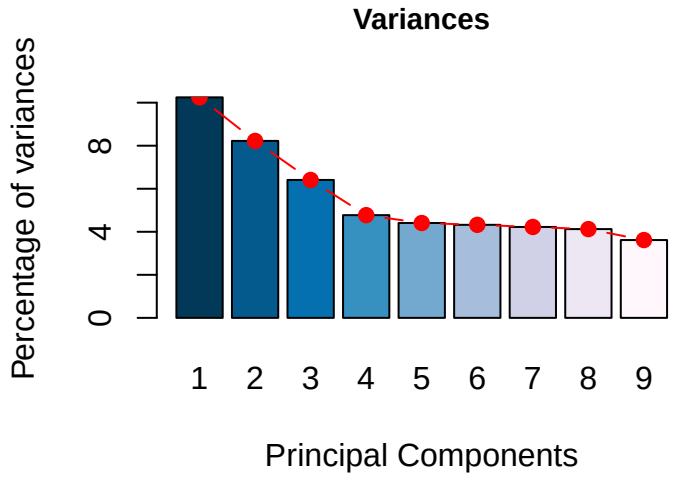

- Естественно высокая корреляция между потреблением спиртного в будние дни и выходные. Если уж выпивать, так основательно.
- Если школьник проводит время вне дома (“goes out” ?), то скорее всего он будет выпивать на выходных. Или наоборот?
- Мужской пол прибавляет количество выпитого на выходных.
- Значительная корреляция между уровнем образования супругов. Понять можно. Муж и жена - одна сатана.
- Средняя оценка положительно коррелирует с желанием поступить в вуз.
- Количество проваленных прошлых курсов + связано с возрастом - вот откуды появились “лбы” - это второгодники и им подобные.

## 3. Предсказание средней оценки школьника.

Попробуем построить модель для предсказания средней оценки школьника, но с единственным ограничением, из-за которого эта модель будет заведомо обречена иметь низкое качество - она должна быть нормально визуализирована.

Признаков слишком много для человеческого глаза. Применим метод главных компонет к датафрейму, который теперь стал стандартизованной (z-scored) матрицей.

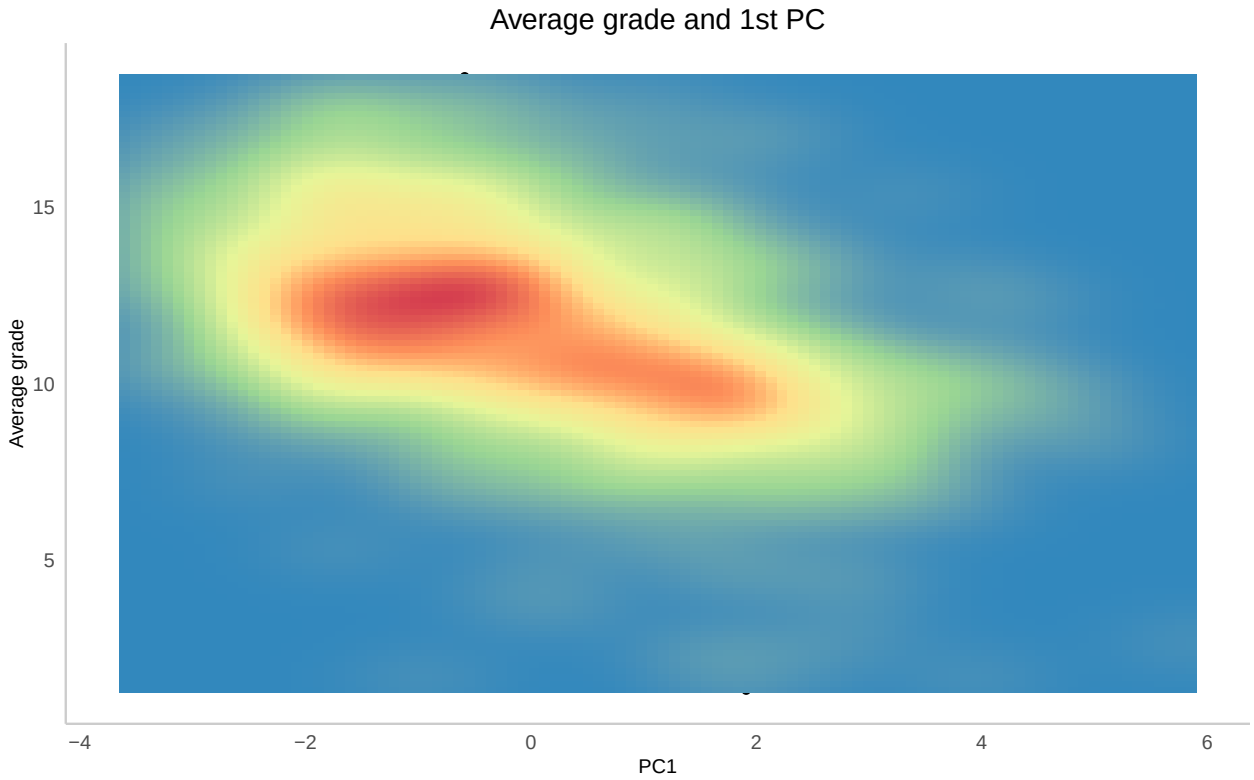
Первые главные компоненты объясняют лишь небольшую часть разброса данных:




По осям - значения первых двух главных компонент. Размер точки - это количество выпитого спиртного на выходных.


Большие “пузыри” (“школьники - алкоголики”) концентрируются в верхней части графика.

Можно попробовать построить линейную регрессию, где объясняющими переменными будут первые главные компоненты. Так как модель должна быть визуализирована (т.е. представлена в 2-х мерном пространстве) - используем лишь первую главную компоненту по оси X. По оси Y - средняя оценка школьника. Как в таком случае будут распределены наблюдения?



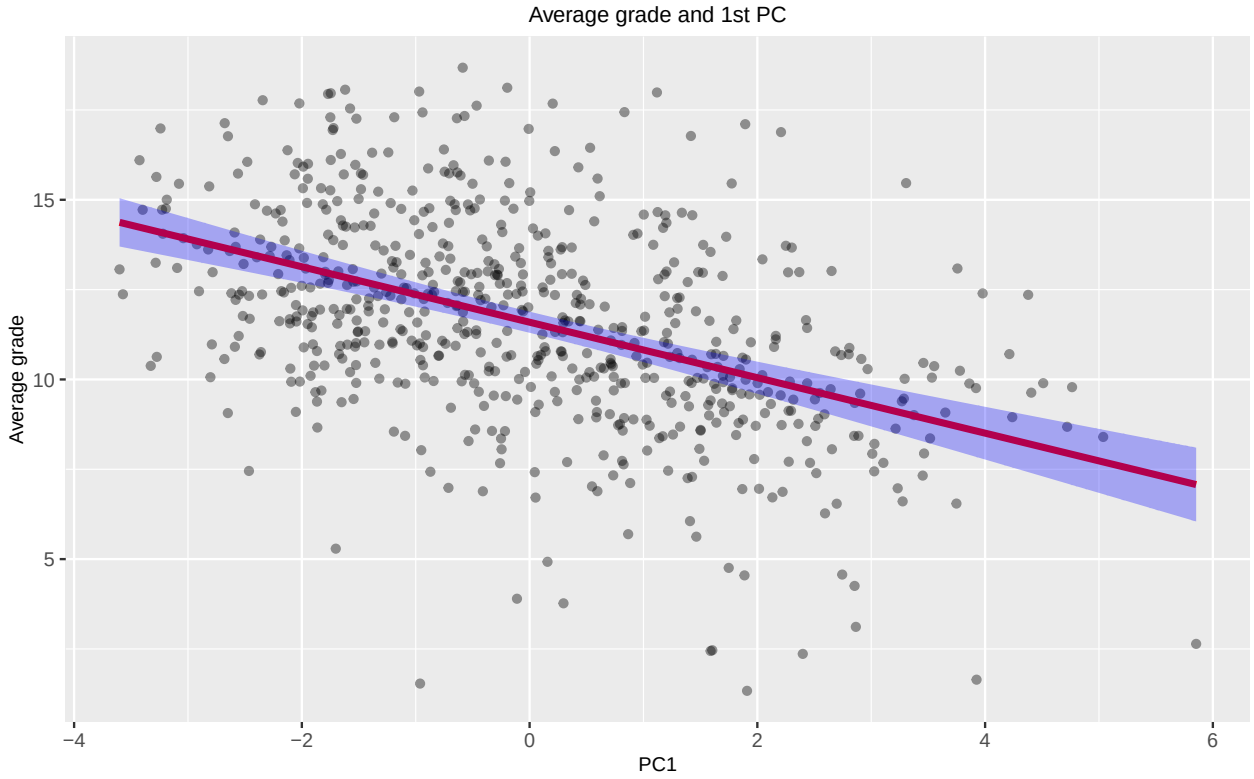

“Облако” имеет небольшую направленность, и линейная регрессия будет иметь отрицательное значение коэффицента наклона:

```
model <- lm(data = df_scores, Aver_grade ~ PC1)
summary(model)
```

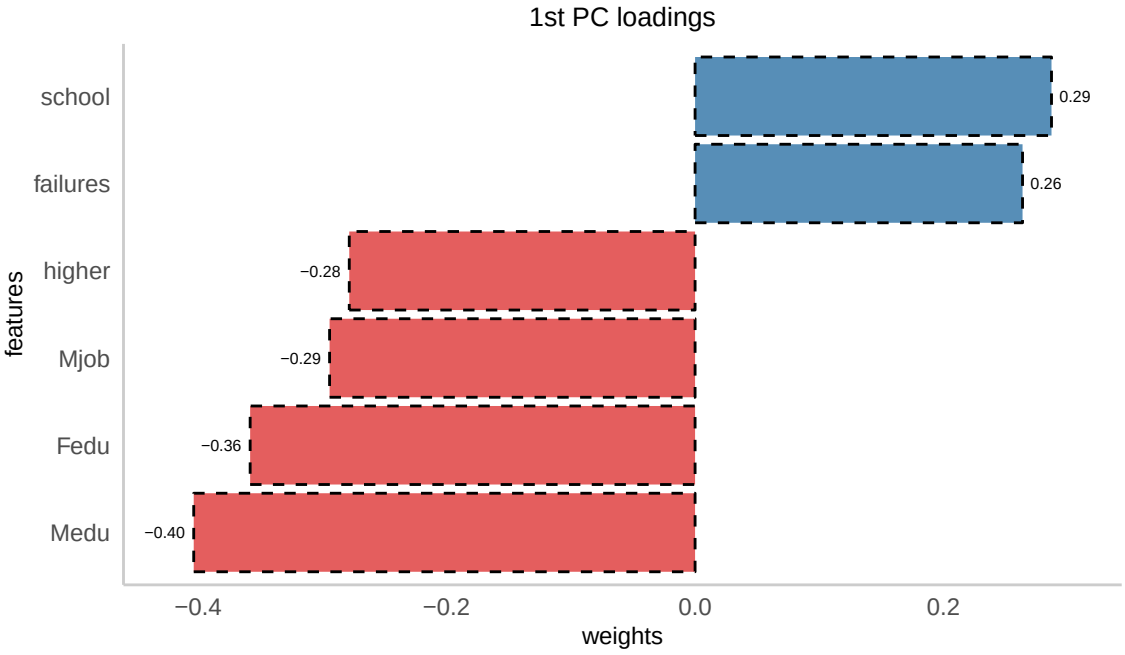
```
##
## Call:
## lm(formula = Aver_grade ~ PC1, data = df_scores)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.668  -1.481  -0.051   1.481   7.271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.59265    0.09809  118.18   <2e-16 ***
## PC1         -0.77184    0.05599  -13.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.524 on 660 degrees of freedom
## Multiple R-squared:  0.2236, Adjusted R-squared:  0.2224
## F-statistic:   190 on 1 and 660 DF,  p-value: < 2.2e-16
```

Коэффициентов всего 2, поэтому они значимые.

Теперь можно нарисовать линию оцененной регрессии вместе со всеми наблюдениями:




Первая главная компонента отрицательно влияет на средний балл ученика. Интересно, из каких признаков в основном состоит эта компонента?



Признаки с положительными весами положительно влияют на первую главную компоненту, которая в свою очередь снижает средний балл школьника согласно оцененной модели. Получается интересная картинка:

- Школа Gabriel Pereira и наличие проблем с прошлыми курсами негативно воздействуют на средний балл школьника.
- Высокий уровень образования родителей, а также занимаемая ими хорошая должность увеличивают средний балл.
- Желание поступить в университет также благоприятствует получению хороших отметок.

# Приложение

Создание корреляционной матрицы:

```
cormat <- round(cor(df_rec), 2)
get_upper_tri <- function(cormat){
  cormat[lower.tri(cormat)]<- NA
  return(cormat)
}
reorder_cormat <- function(cormat){
  dd <- as.dist((1-cormat)/2)
  hc <- hclust(dd)
  cormat <-cormat[hc$order, hc$order]
}
cormat <- reorder_cormat(cormat)
upper_tri <- get_upper_tri(cormat)
melted_cormat <- melt(upper_tri, na.rm = TRUE)

ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value)) +
            geom_tile(color = "white") + scale_fill_gradient2(low = "blue", high = "red", mid = "white",
            midpoint = 0, limit = c(-1,1), space = "Lab",
            name = "Pearson\nCorrelation") +
            theme_minimal()+ # minimal theme
            theme(axis.text.x = element_text(angle = 45, vjust = 1, size = 12, hjust = 1)) + coord_fixed()

ggheatmap <- ggheatmap +
             theme(
             axis.title.x = element_blank(),
             axis.title.y = element_blank(),
             axis.text.x = element_text(size = 9, angle = 90),
             axis.text.y = element_text(size = 9),
             panel.grid.major = element_blank(),
             panel.border = element_blank(),
             panel.background = element_blank(),
             axis.ticks = element_blank(),
             legend.justification = c(1, 0),
             legend.position = c(0.6, 0.7),
             legend.direction = "horizontal")+
             guides(fill = guide_colorbar(barwidth = 7, barheight = 1, title.position = "top", title.hjust = 0.5)) +
             geom_text(aes(Var2, Var1, label = value), color = "black", size = 2.5)
```

Метод главных компонент и визуализация результатов:

```
# Apply PCA approach for visualization and simple regression:
pca <- prcomp(subset(df_rec, select = -c(aver_grade)), scale = TRUE, retx = TRUE)
eigvalues <- (pca$sdev)^2
variance <- eigvalues * 100 / sum(eigvalues)
cumvar <- cumsum(variance)
df_pca <- data.frame(eigvalues = eigvalues, variance = variance, cumvariance = cumvar)

# Show how much variance are explained by first several principal components:
bp_pca <- barplot(df_pca[1:9, 2], names.arg = 1:9, cex.main = 0.9,
        main = "Variances",
```

```
        xlab = "Principal Components",
        ylab = "Percentage of variances",
        col = rev(brewer.pal(9, 'PuBu')))

lines(x = bp_pca, y = df_pca[1:9, 2], type = 'b', pch = 19, col = "red")

df_scores <- data.frame(PC1 = pca$x[, 1], PC2 = pca$x[, 2], Walc = df_rec$Walc, Aver_grade = df_rec$aver_grade)
scatter_pca <- ggplot(df_scores, aes(PC1, PC2)) +
            geom_jitter(alpha = 0.7, aes(colour = factor(Walc), size = factor(Walc))) +
            ggtitle("Weekend alcohol consumption") +
            theme(plot.title = element_text(size = 10),
                  axis.title.x = element_text(size = 9),
                  axis.title.y = element_text(size = 9),
                  legend.title = element_text(size = 9))
```

Оценка и визуализация простой линейной регрессии:

```
# Model regression fit of average grade against 1st PC.
# Firstly take a look at scatter plot and its density:
cont_plot <- ggplot(data = df_scores, aes(x = PC1, y = Aver_grade)) +
          geom_point() +
          ylab('Average grade') + ggtitle('Average grade and 1st PC') +
          stat_density_2d(geom = "raster", aes(fill = ..density..), contour = FALSE) +
          scale_fill_distiller(palette = "Spectral") +
          theme(panel.background = element_blank(),
              panel.grid.minor = element_blank(),
              axis.ticks  = element_blank(),
              axis.line   = element_line(colour=NA),
              axis.line.x = element_line(colour="grey80"),
              axis.title.x = element_text(size = 9),
              axis.title.y = element_text(size = 9),
              legend.title = element_text(size = 9),
              axis.line.y = element_line(colour="grey80")) +
         guides(fill = FALSE)

model <- lm(data = df_scores, Aver_grade ~ PC1)
summary(model)

regr_plot <- ggplot() + geom_jitter(data = df_scores, aes(x = PC1, y = Aver_grade), alpha = 0.4) +
  ylab('Average grade') + ggtitle('Average grade and 1st PC')
PC1_min <- min(df_scores$PC1)
PC1_max <- max(df_scores$PC1)
grade_pred <- data_frame(PC1 = seq(from = PC1_min, to = PC1_max, length.out = length(df_scores$PC1)))
grade_pred <- augment(model, newdata = grade_pred)
regr_plot <- regr_plot + geom_line(data = grade_pred, color = "red",
                                   aes(x = grade_pred$PC1, y = grade_pred$.fitted), size = 1.5)
grade_pred <- mutate(grade_pred, left = .fitted - 3 * .se.fit, right = .fitted + 3 * .se.fit)
regr_plot <- regr_plot + geom_ribbon(data = grade_pred, fill = 'blue',
          aes(x = grade_pred$PC1, ymin = grade_pred$left, ymax = grade_pred$right), alpha = .3) +
          theme(plot.title = element_text(size = 10),
          axis.title.x = element_text(size = 9),
          axis.title.y = element_text(size = 9),
          legend.title = element_text(size = 9))
```

Составляющие первой главной компоненты:

```
# Explain which features mostly affect the 1st PC:
df_loadings <- data.frame(PC1 = sort(pca$rotation[abs(pca$rotation[, 1]) > 0.25, 1]))
bar_loadings <- ggplot(df_loadings, aes(x = rownames(df_loadings), y = PC1)) +
                geom_bar(position="identity", stat="identity", fill=ifelse(df_loadings$PC1 > 0,
                         rgb(45,114,166, maxColorValue = 255),
                         rgb(222,54,54, maxColorValue=255)), alpha = 0.8, color = 'black', linetype = 'dashed') +
                ggtitle("1st PC loadings") + xlab('features') + ylab('weights') +
                geom_text(aes(x = rownames(df_loadings),
                          y = PC1 + 0.02 * sign(PC1),
                          label=format(PC1, digits=2)),
                          hjust=0.6,
                          size=2.,
                          color=rgb(0,0,0, maxColorValue=255)) +
                theme(panel.background = element_blank(),
                      panel.grid.minor = element_blank(),
                      axis.ticks  = element_blank(),
                      axis.line   = element_line(colour=NA),
                      axis.line.x = element_line(colour="grey80"),
                      axis.line.y = element_line(colour="grey80"),
                      plot.title = element_text(size = 10),
                      axis.title.x = element_text(size = 9),
                      axis.title.y = element_text(size = 9)) +
                coord_flip() +
                scale_x_discrete(limits=c(rownames(df_loadings)), labels = c(rownames(df_loadings)))

bar_loadings
```