Research project: Alcohol Consumption between Students

Contents

Team members: Alina Voronina, Anastasiia Havryliv

Goal of the project

We want to analyze the factors that influence alcohol consumption between students, and see dependence and/or correlation between them.

Data

We use a dataset from kaggle (student-mat.csv) available by the following link: https://www.kaggle.com/uciml/student-alcohol-consumption

Remarks

In the report, we will use the following notions:

- Dalc - level of students' alcohol consumption on workdays
- Walc - level of students' alcohol consumption on weekends

All other notions are explained in the database context by the above link.

- student-mat.csv dataset - used for training
- student-por.csv dataset - used once for testing the linear regression model

Packages that are to be used:

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
library(ggplot2)
```

Data processing

We substitute all non-numeric values with numeric ones, because that will facilitate the data processing later. We also drop some columns (as school name, address, etc) that complicate the analysis and do not allow to precisely process other information (it is almost impossible to find correlations between unprocessed strings and numeric values).

dropped columns: school (1), address (4), mjob (9), fjob(10), reason (11), guardian (12)

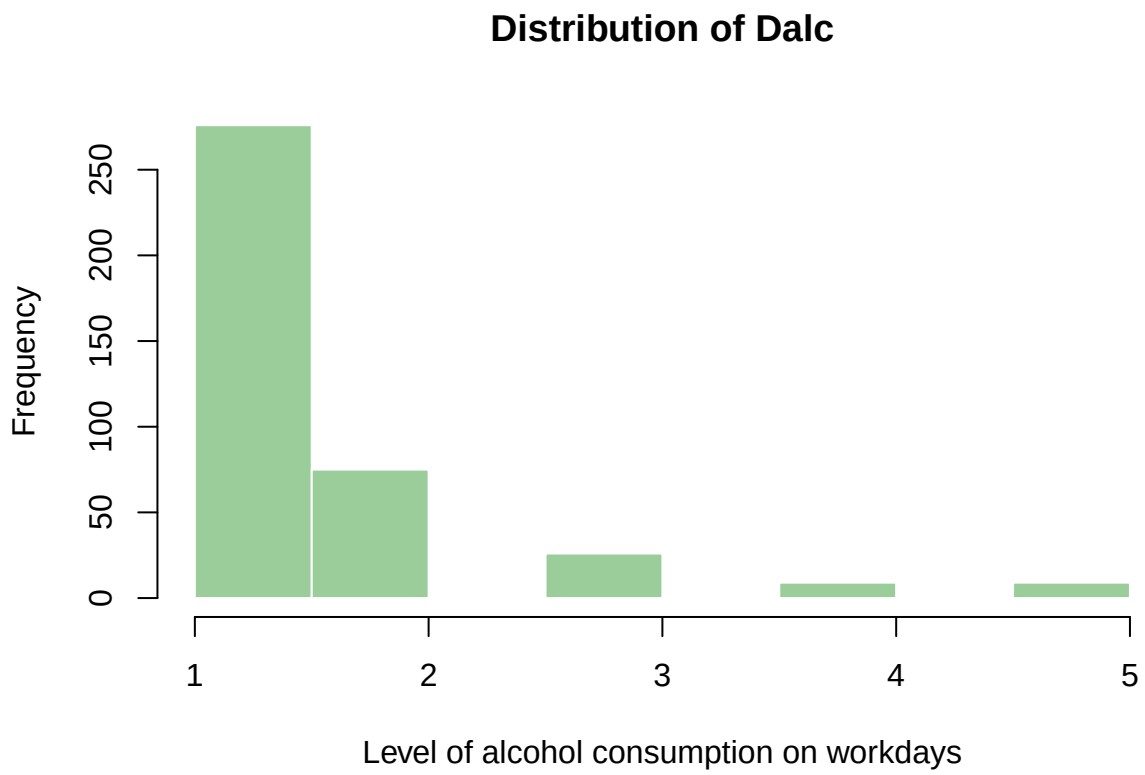
```
data = read.csv("student-mat.csv")

data = data[-c(1, 4, 9, 10, 11, 12)]

data$sex = ifelse(data$sex == "M", as.integer(0), as.integer(1))
data$famsize = ifelse(data$famsize == "LE3", as.integer(0), as.integer(1))
data$Pstatus = ifelse(data$Pstatus == "A", as.integer(0), as.integer(1))

data$schoolsup = ifelse(data$schoolsup == "no", as.integer(0), as.integer(1))
data$famsup = ifelse(data$famsup == "no", as.integer(0), as.integer(1))
data$paid = ifelse(data$paid == "no", as.integer(0), as.integer(1))
data$activities = ifelse(data$activities == "no", as.integer(0), as.integer(1))
data$nursery = ifelse(data$nursery == "no", as.integer(0), as.integer(1))
data$higher = ifelse(data$higher == "no", as.integer(0), as.integer(1))
data$internet = ifelse(data$internet == "no", as.integer(0), as.integer(1))
data$romantic = ifelse(data$romantic == "no", as.integer(0), as.integer(1))

head(data)
```

```
##   sex age famsize Pstatus Medu Fedu traveltime studytime failures schoolsup
## 1   1  18       1       0    4    4          2         2        0         1
## 2   1  17       1       1    1    1          1         2        0         0
## 3   1  15       0       1    1    1          1         2        3         1
## 4   1  15       1       1    4    2          1         3        0         0
## 5   1  16       1       1    3    3          1         2        0         0
## 6   0  16       0       1    4    3          1         2        0         0
##   famsup paid activities nursery higher internet romantic famrel freetime goout
## 1      0    0          0       1      1        0        0      4        3     4
## 2      1    0          0       0      1        1        0      5        3     3
## 3      0    1          0       1      1        1        0      4        3     2
## 4      1    1          1       1      1        1        1      3        2     2
## 5      1    1          0       1      1        0        0      4        3     2
## 6      1    1          1       1      1        1        0      5        4     2
##   Dalc Walc health absences G1 G2 G3
## 1    1    1      3        6  5  6  6
```

```
## 2    1    1       3        4  5  5  6
## 3    2    3       3       10  7  8 10
## 4    1    1       5        2 15 14 15
## 5    1    2       5        4  6 10 10
## 6    1    2       5       10 15 15 15
```

Let's look at the Dalc and Walc distribution.

```
hist(data$Dalc, col="darkseagreen3", border="white", main="Distribution of Dalc", xlab = "Level of alcol
```




```
hist(data$Walc, col="darkseagreen3", border="white", main="Distribution of Walc", xlab = "Level of alcol
```

Conclusion: students drink way too much on weekends!

Finding correlations

Now we want to find correlation between Dalc/Walc and other factors.

```
correlation = cor(data)

cat("Correlation on workdays\n\n")
```

```
## Correlation on workdays
```

```
print(sort(abs(correlation[, "Dalc"])))
```

```
##        Fedu    romantic        Medu   schoolsup     Pstatus      famsup
## 0.002386429 0.015120705 0.019834099 0.021485100 0.030589889 0.031575204
##    internet          G3        paid          G2  activities      higher
## 0.036210377 0.054660041 0.062465362 0.064120183 0.066508094 0.069828063
##      health      famrel     nursery          G1     famsize    absences
## 0.077179582 0.077594357 0.084848638 0.094158792 0.101521261 0.111908026
##         age    failures  traveltime   studytime    freetime       goout
## 0.131124605 0.136046931 0.138325309 0.196019263 0.209000848 0.266993848
##         sex        Walc        Dalc
## 0.268170983 0.647544230 1.000000000
```

```
cat("\n\n")
```

```
cat("Correlation on weekends\n\n")
```

```
## Correlation on weekends
```

```
print(sort(abs(correlation[, "Walc"])))
```

```
##    Pstatus   romantic   internet       Fedu activities       Medu         G3
## 0.00604478 0.01014095 0.01168720 0.01263102 0.03747670 0.04712346 0.05193932
##       paid         G2     famsup  schoolsup     health    nursery     higher
## 0.06045364 0.08492735 0.08668793 0.08715174 0.09247632 0.09953353 0.10033961
##    famsize     famrel        age         G1 traveltime   absences   failures
## 0.10342501 0.11339731 0.11727605 0.12617921 0.13411575 0.13629110 0.14196203
##   freetime  studytime        sex      goout       Dalc       Walc
## 0.14782181 0.25378473 0.27419377 0.42038575 0.64754423 1.00000000
```

As we can see, Walc and Dalc mostly correlate with sex, goout, studytime, and freetime.

- Dalc also correlates with traveltime, failures and age.
- Walc also correlates with failures, absences and traveltime.

We can also spot some interesting facts: - On workdays your gender has more importance than the frequency you go out, meanwhile on weekends the more you go out the more likely you are to drink :)

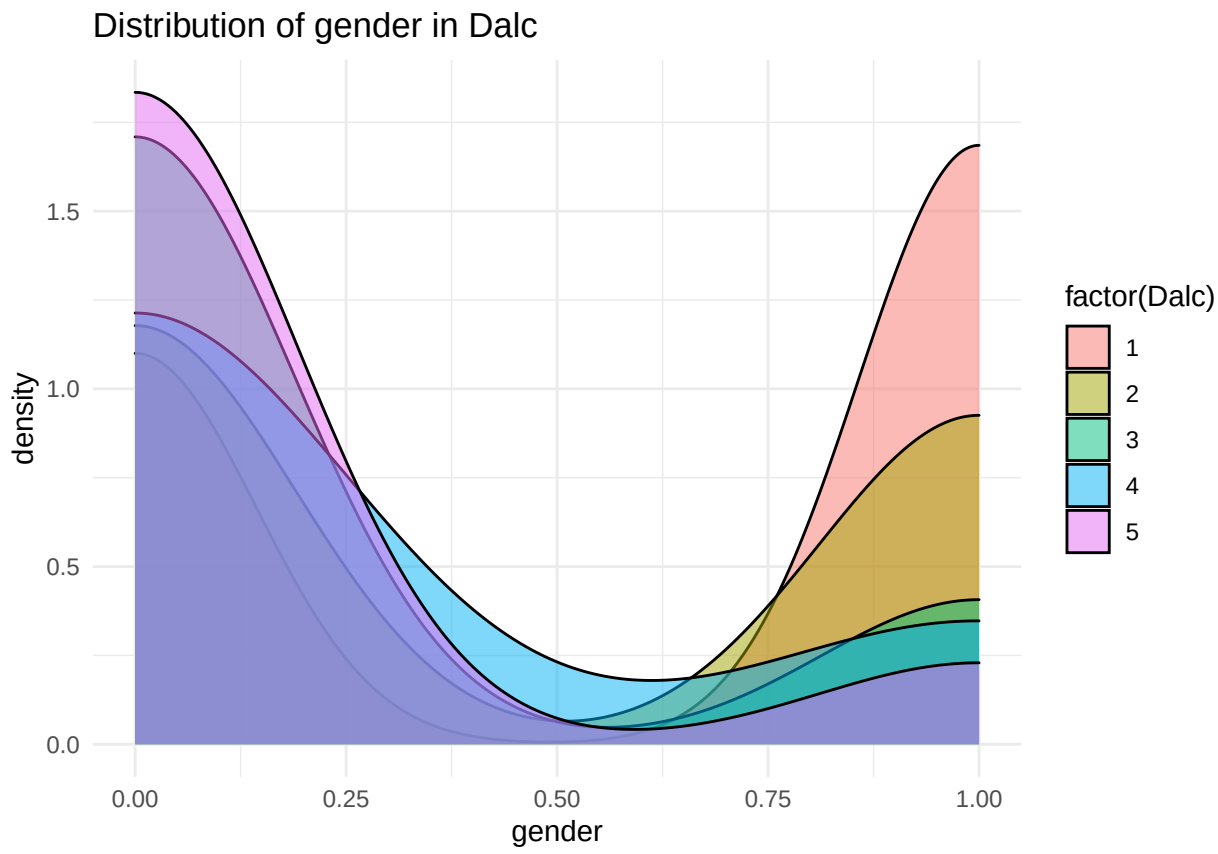
- On workdays students tend to get drunk because of having more freetime, meanwhile on weekends - because of studying too much :)

Surpisingly, parents' status, education and other family factors show so little correlation.

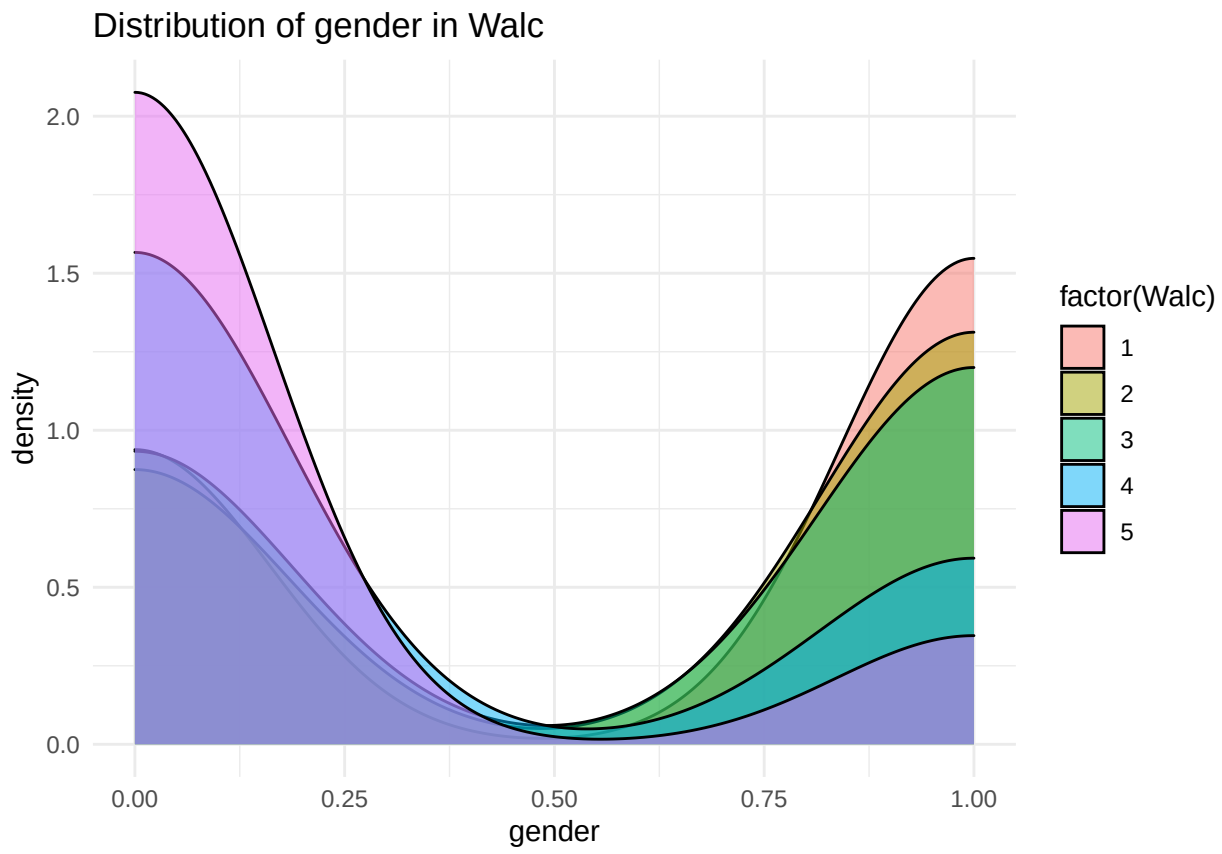
Displaying the influence of different factors

Gender

```
ggplot(data,aes(x=sex, fill=factor(Dalc))) + geom_density(alpha=0.5)+
 xlab(label = "gender")+
 ggtitle("Distribution of gender in Dalc")+
 theme_minimal()
```



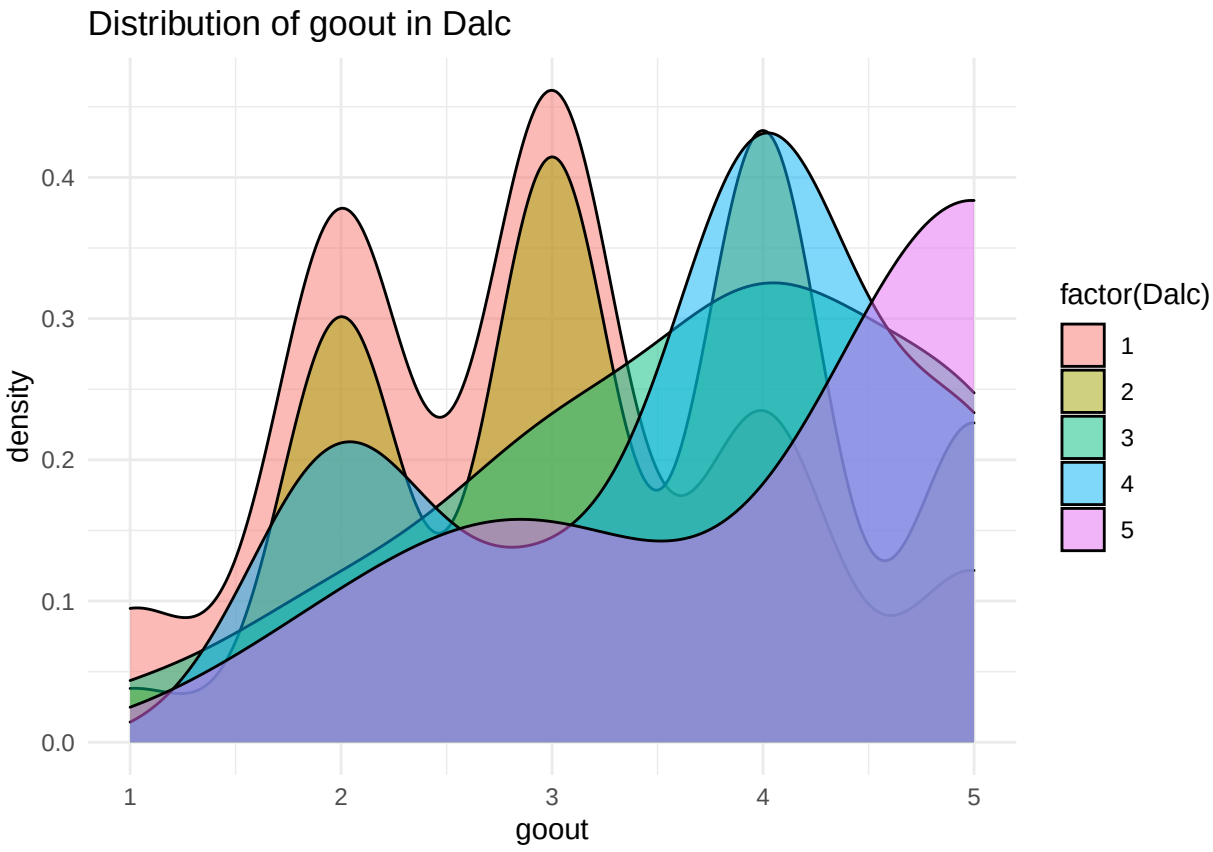

```
ggplot(data,aes(x=sex, fill=factor(Walc))) + geom_density(alpha=0.5)+
 xlab(label = "gender")+
 ggtitle("Distribution of gender in Walc")+
 theme_minimal()
```



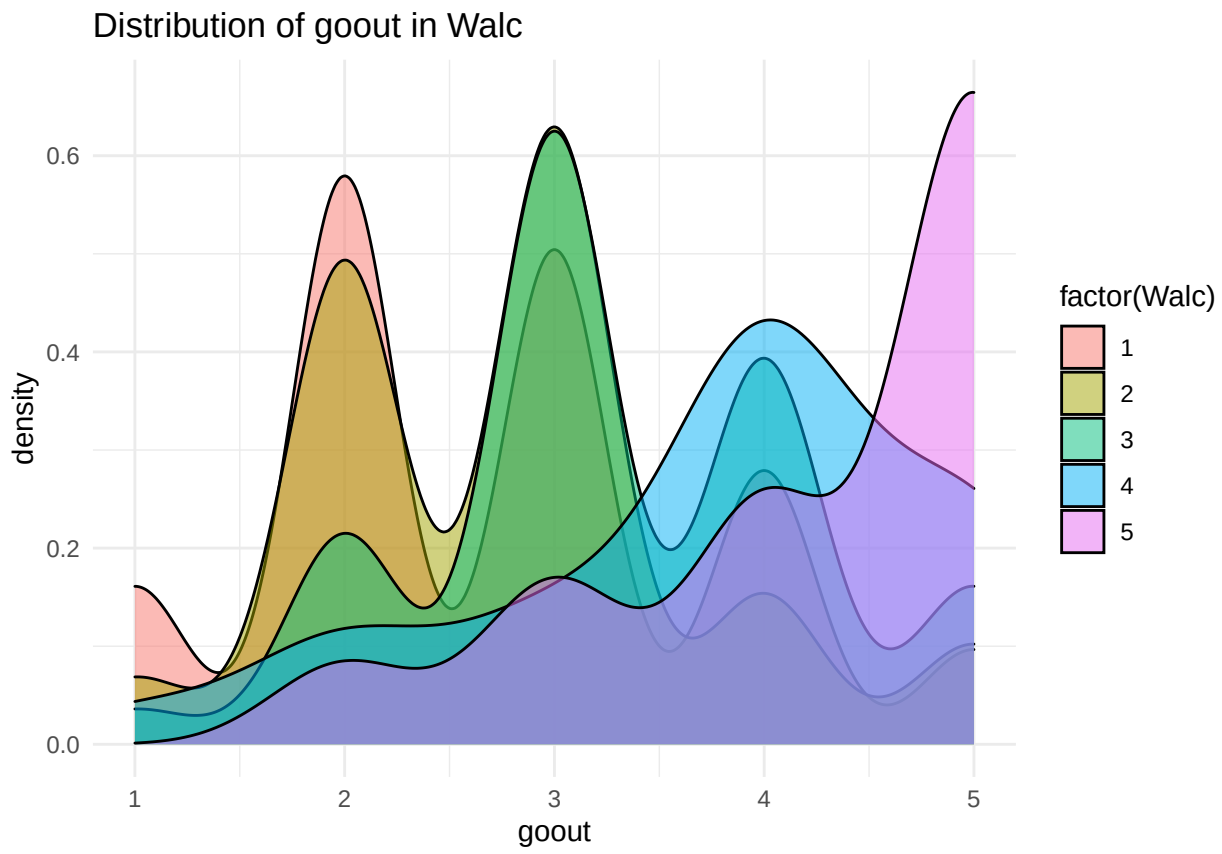
Remembering that 0 is for male, 1 is for female, we can see, that women do not get as drunk as man. It can be notices that on weekends both men and women drink way more than on workdays.

Going out

```
ggplot(data,aes(x=goout, fill=factor(Dalc))) + geom_density(alpha=0.5)+
 xlab(label = "goout")+
 ggtitle("Distribution of goout in Dalc")+
 theme_minimal()
```



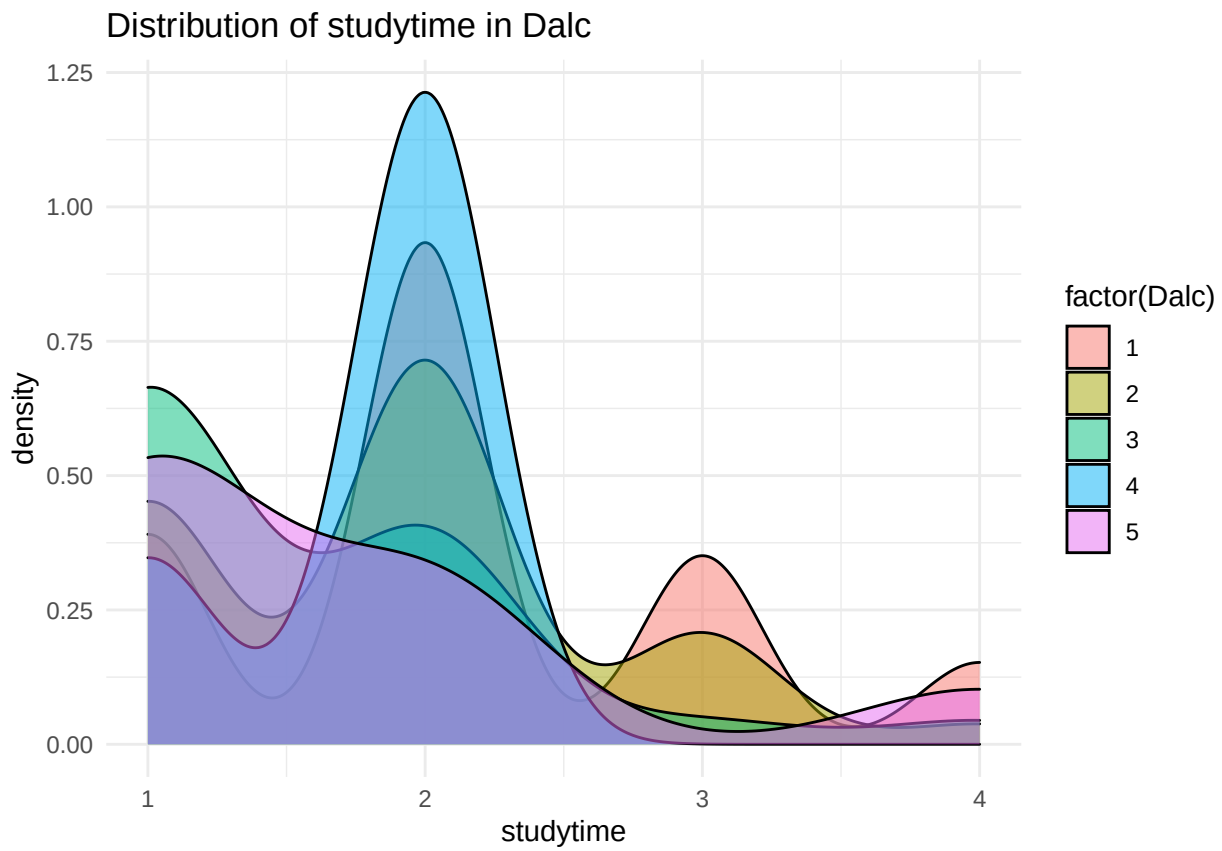

```
ggplot(data,aes(x=goout, fill=factor(Walc))) + geom_density(alpha=0.5)+
 xlab(label = "goout")+
 ggtitle("Distribution of goout in Walc")+
 theme_minimal()
```



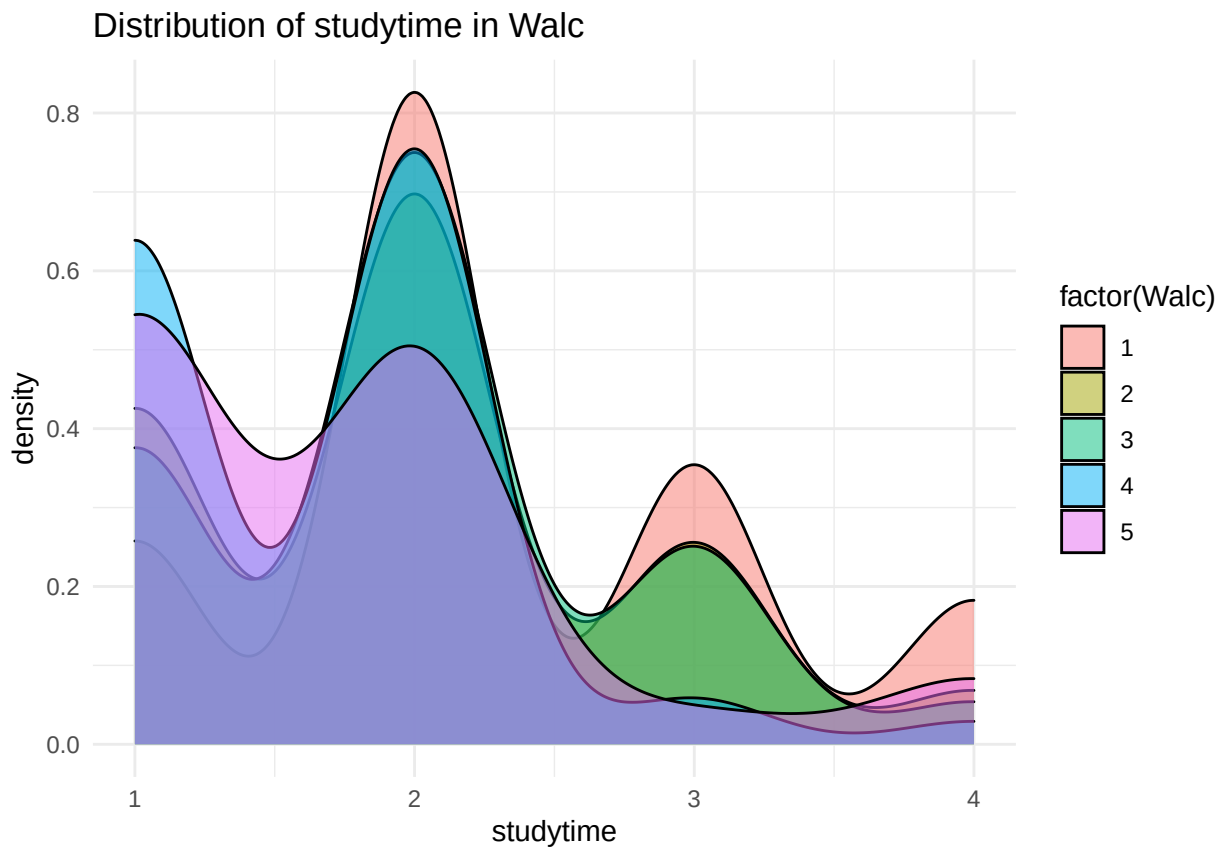
Again, on weekends the level of alcohol consumption is skyrocketing.

Studytime

```
ggplot(data,aes(x=studytime, fill=factor(Dalc))) + geom_density(alpha=0.5)+
 xlab(label = "studytime")+
 ggtitle("Distribution of studytime in Dalc")+
 theme_minimal()
```



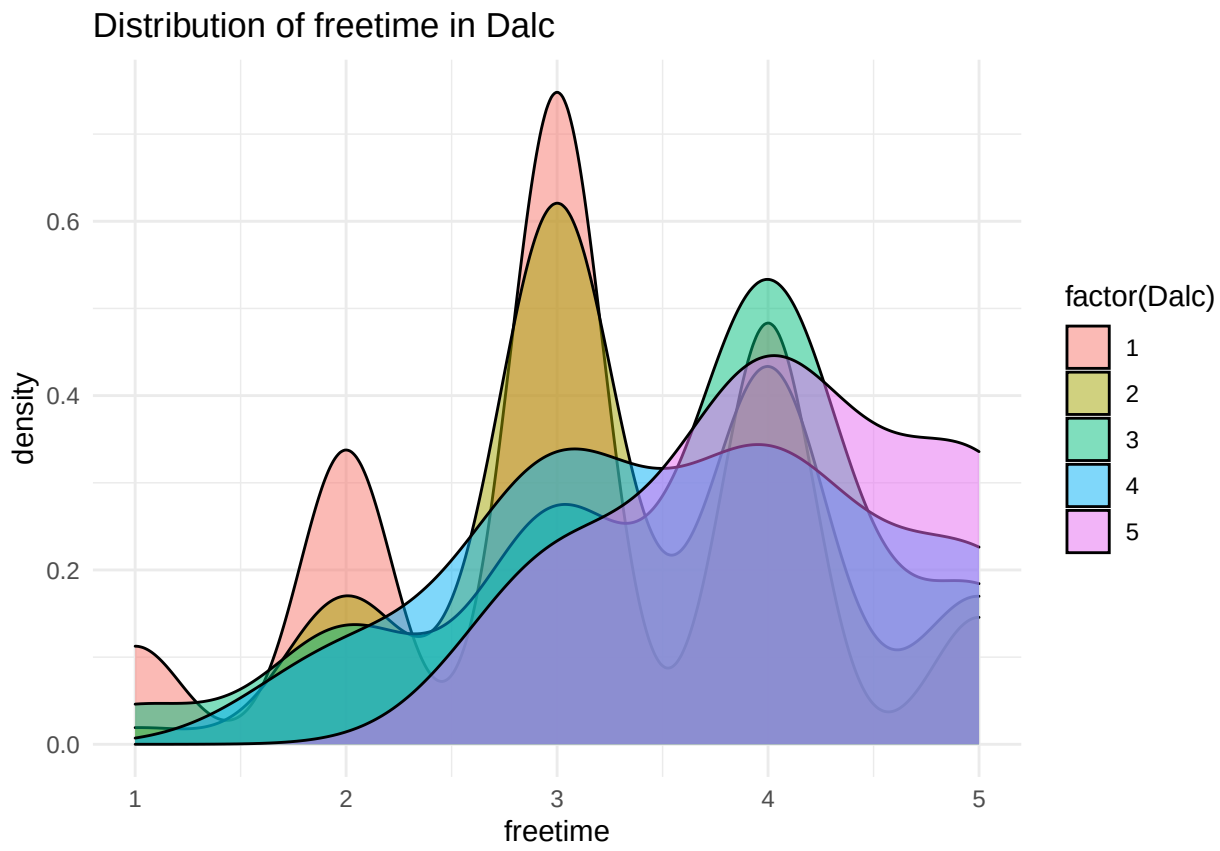

```
ggplot(data,aes(x=studytime, fill=factor(Walc))) + geom_density(alpha=0.5)+
 xlab(label = "studytime")+
 ggtitle("Distribution of studytime in Walc")+
 theme_minimal()
```



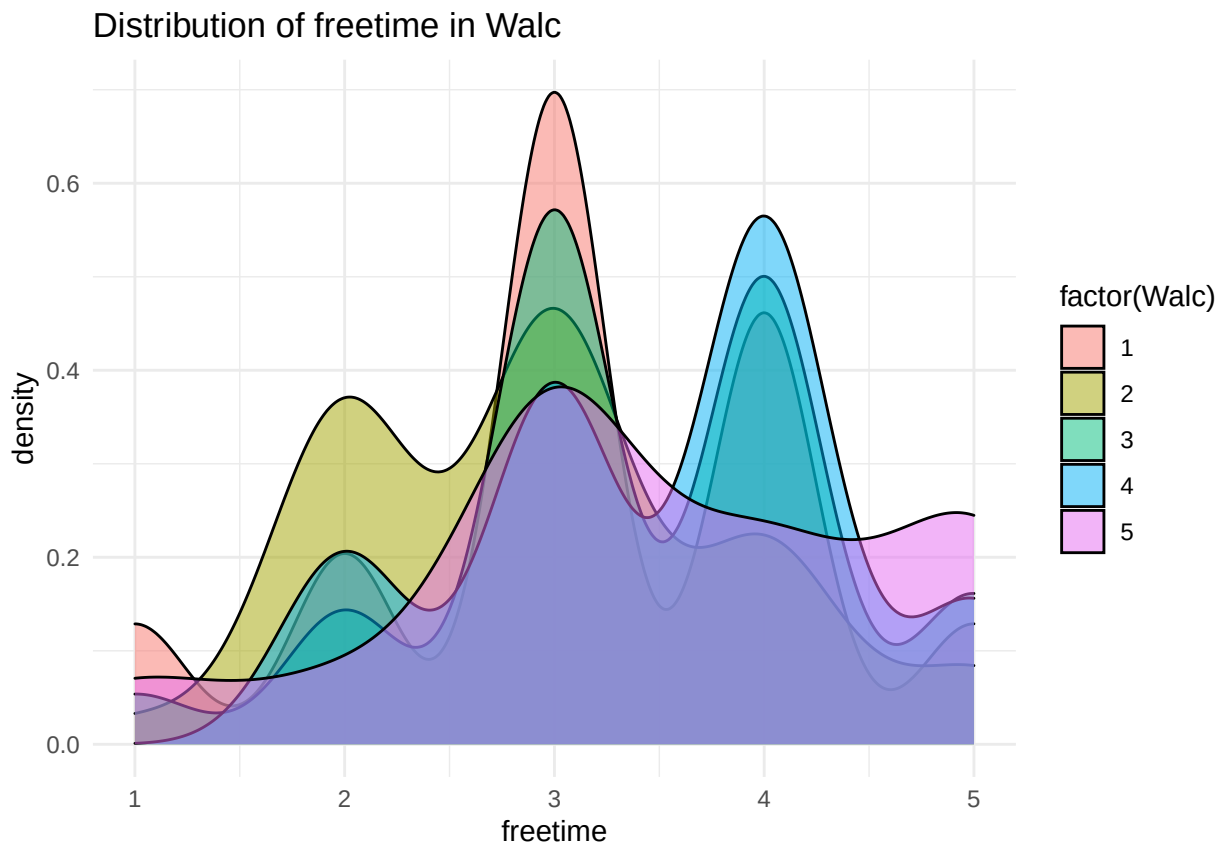
People who study more, have higher level of alcohol consumption on workdays (no jokes about programmers constantly drinking beer).

Freetime

```
ggplot(data,aes(x=freetime, fill=factor(Dalc))) + geom_density(alpha=0.5)+
 xlab(label = "freetime")+
 ggtitle("Distribution of freetime in Dalc")+
 theme_minimal()
```




```
ggplot(data,aes(x=freetime, fill=factor(Walc))) + geom_density(alpha=0.5)+
 xlab(label = "freetime")+
 ggtitle("Distribution of freetime in Walc")+
 theme_minimal()
```



The plots are very similar. We can see that on the weekends the level of alcohol consumption rises a bit.

Testing the hypothesis

Dependence on studytime

Taking the above plots into consideration, it would be interesting to test the dependence of studytime and Dalc/Walc.

Firstly, let's introduce the following hypothesis:

H_0: Level of alcohol consumption does not depend on studytime. H_1: Level of alcohol consumption depends on studytime.

As long as we need to test whether the two characteristics are (in)dependent, we will use the chi-squared test for independence.

The value of the test statistics in chi-square test is the next: $\chi^2 = \sum_{i=1}^{rows} \sum_{j=1}^{columns} \frac{o_{ij}-e_{ij}}{e_{ij}}$

where oij is the observed value, e_{ij} is the predicted/expected value.

Under H_0 this statistics has approximate distribution χ^2_r with r degrees of freedom. In this case, we have $(rows-1)(columns-1)$degrees of freedom. Thus, we need to compare the above statistics with $\chi^2_{(rows-1)(columns-1)}$, calculate the p-value and make conclusions based on that.

To perform chi-squared test, we need to form a dataframe first. Rows are in charge of studytime, columns are in charge of Dalc.

```
studytimes = sort(unique(data$studytime))
dalcdf = data.frame(matrix(ncol = 5, nrow = length(studytimes)))
dalcdf[is.na(dalcdf)] = 0

colnames(dalcdf) = c('1', '2', '3', '4','5')
rownames(dalcdf) = studytimes

for (row in 1:nrow(data)) {
  dlevel = data[row, "Dalc"]
  stime = data[row, "studytime"]
  dalcdf[stime, dlevel] = dalcdf[stime, dlevel] + 1
}
```

Then we will apply the chi-square test to the obtained dataframe.

```
chisq = chisq.test(dalcdf)
```

```
## Warning in chisq.test(dalcdf): Chi-squared approximation may be incorrect
```

```
print(chisq)
```

```
##
##  Pearson's Chi-squared test
##
## data:  dalcdf
## X-squared = 30.091, df = 12, p-value = 0.002706
```

We can notice that the p-value is small enough to reject the null hypothesis. Therefore, we can assume that the Dalc level is dependent on the studytime.

Now let's perform the same test, but to find whether Walc is dependent on studytime.

```
studytimes = sort(unique(data$studytime))
walcdf = data.frame(matrix(ncol = 5, nrow = length(studytimes)))
walcdf[is.na(walcdf)] = 0

colnames(walcdf) = c('1', '2', '3', '4','5')
rownames(walcdf) = studytimes

for (row in 1:nrow(data)) {
  dlevel = data[row, "Walc"]
  stime = data[row, "studytime"]
  walcdf[stime, dlevel] = walcdf[stime, dlevel] + 1
}
```

Performing the chi-square test.

```
chisq = chisq.test(walcdf)
```

```
## Warning in chisq.test(walcdf): Chi-squared approximation may be incorrect
```

```
print(chisq)
```

```
##
##  Pearson's Chi-squared test
##
## data:  walcdf
## X-squared = 34.902, df = 12, p-value = 0.0004854
```

We see that the p-value is even smaller than in the previous case with Dalc. That means that we reject the null hypothesis again, meaning Walc depends on the studytime.

Conclusions

Level of students' alcohol consumption on workdays (Dalc) depends on studytime.

Level of students' alcohol consumption on weekends (Walc) depends on studytime.

Linear regression

We will train the model on students-mat.csv dataset (the one we used everywhere before), and test the model on students-por.csv dataset (available by the same link in the beginning).

Firstly, a little data processing.

```
test = read.csv("student-por.csv")
test = test[-c(1, 4, 9, 10, 11, 12)]
test$sex = ifelse(test$sex == "M", as.integer(0), as.integer(1))
```

Let's now perform linear regression by developing a model.

In this sections we create a model for further prediction of workday alcohol consumption, based on the following factors: sex, go out time, study time, and free time.

And we can see the summary with all the estimators for this model, which will help with predictions.

```
model = lm(Dalc ~ sex + goout + studytime + freetime, data=data)
summary(model)
```

```
##
## Call:
## lm(formula = Dalc ~ sex + goout + studytime + freetime, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1785 -0.5324 -0.2074  0.2562  3.5978
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.11312    0.20911   5.323 1.72e-07 ***
## sex         -0.35293    0.08926  -3.954 9.13e-05 ***
## goout        0.17708    0.03896   4.545 7.32e-06 ***
## studytime   -0.11604    0.05218  -2.224   0.0267 *
```

```
## freetime     0.07399    0.04467   1.656   0.0985 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8249 on 390 degrees of freedom
## Multiple R-squared:  0.151,  Adjusted R-squared:  0.1423
## F-statistic: 17.34 on 4 and 390 DF,  p-value: 4.189e-13
```

Here we test the accuracy of our model, and we see that it is not very close to 1, and it is because we don't take some factors into consideration (because they have very low impact and make our calculations more complicated).

```
pred <- c(as.integer(predict(model, newdata = test)))
count = sum(c(ifelse(test$Dalc == pred, 1, 0)))
accuracy = count/nrow(test)
accuracy
```

```
## [1] 0.6101695
```

Here we start with the predictions for workday alcohol consumption. We assume that user inputs some data, then we create a single-row database, which represents this input. With the help of the model we predict the level of Dalc that the user has.

For example, let's assume user entered the next data:

- sex: 1 (F)
- studytime: 2
- goout: 1
- freetime: 5

```
# Data input started
sex <- as.integer(1)
studytime <- as.integer(2)
goout <- as.integer(1)
freetime <- as.integer(5)
# Data input ended

newdata = data.frame(sex, studytime, goout, freetime)

prediction <- predict(model, newdata = newdata)

if (prediction < 0){
  prediction <- 0
}

if (prediction > 5){
  prediction <- 5
}

cat("Prediction of weekday alcohol consumption:", prediction)
```

```
## Prediction of weekday alcohol consumption: 1.075161
```

The same manipulations can be done to predict the Walc level.

Conclusions

We displayed the correlation between different factors and Dalc/Walc. Having plotted some graphs, we indeed saw those correlations, and assumed that the most significant factors were sex, goout, studytime and freetime.

Our aim was to test the dependency of Dalc/Walc on the above factors. We decided to take studytime factor for that, as we thought that it will be the most interesting to see the dependency of the time student spends studying on their level of alcohol consumption. As a result, we saw that Dalc/Walc indeed depend on studytime.

We implemented a linear regression model to predict the alcohol consumption level. We trained it on students-mat.csv dataset and tested on the students-por.csv one. The model turned out to be approximately 60% accurate. The reason for that is that we took only most correlated features (which we estimated at the beginning) into consideration, and even those features did not show correlation, very close to +-1. We also did not consider all 30+ factors. Moreover, the model predicts a float value in range from 1 to 5, and the actual value is an integer, so we needed to round the result - that also decreased the accuracy in some way. Thus, we have such a score.

We also implemented the prediction for user's data. By changing the values, the user can see their alcohol consumption level.

Happy holidays! :)