Does Alcohol Affects Grades?

Semar Augusto da Cunha Mello Martins

1. Introdução:

O álcool é uma das drogas recreacionais mais comumente usadas pelo mundo, isso acontece devido à sua legalidade e à facilidade de acesso a ela. Contudo, já foi descoberto que o abuso de álcool pode causar vício, o que pode causar múltiplos problemas na vida das pessoas, como falta de motivação para estudar, perca do emprego, entre outros.

Nesse trabalho vamos avaliar se o consumo de álcool tem relação com a nota dos alunos. É importante notar que boa parte dos sujeitos avaliados na base de dados possui menos de 18 anos, por causa disso, a taxa de alcoólatras no estudo será inferior à média mundial. Por esse motivo, não vamos criar um modelo que prevê as notas dos alunos dado uma base de treino, vamos criar um modelo e descobrir qual o peso de cada feature no modelo.

2. Base de Dados:

A base de dados escolhida foi “Student Alcohol Consumption”, disponibilizada inicialmente pelo UCI - Machine Learning Repository mas atualmente só foi possível encontrá-la no Kaggle.

A base corresponde a uma pesquisa com as notas de alunos do ensino médio com notas somente das matérias matemática e português. Além das notas dos alunos, existem vários atributos relacionados à vida de cada aluno.

Features:

- Escola do aluno
- Sexo
- Idade
- Endereço
- Tamanho da família
- Se os pais são separados ou não
- Educação da mãe
- Educação do pai
- Trabalho da mãe
- Trabalho do pai
- Motivo pela escolha da escola
- Quem é o guardião do estudante
- Tempo que o aluno demora para chegar à escola
- Tempo de estudo por semana
- Número de matérias repetidas pelo aluno
- Se a escola dá suporte extra ao aluno
- Se a família suporta o estudo do aluno
- Se o aluno frequenta aulas particulares
- Atividades extra-curriculares
- Se o aluno apresenta interesse em fazer faculdade
- Possui internet?
- Tem relações romanticas?

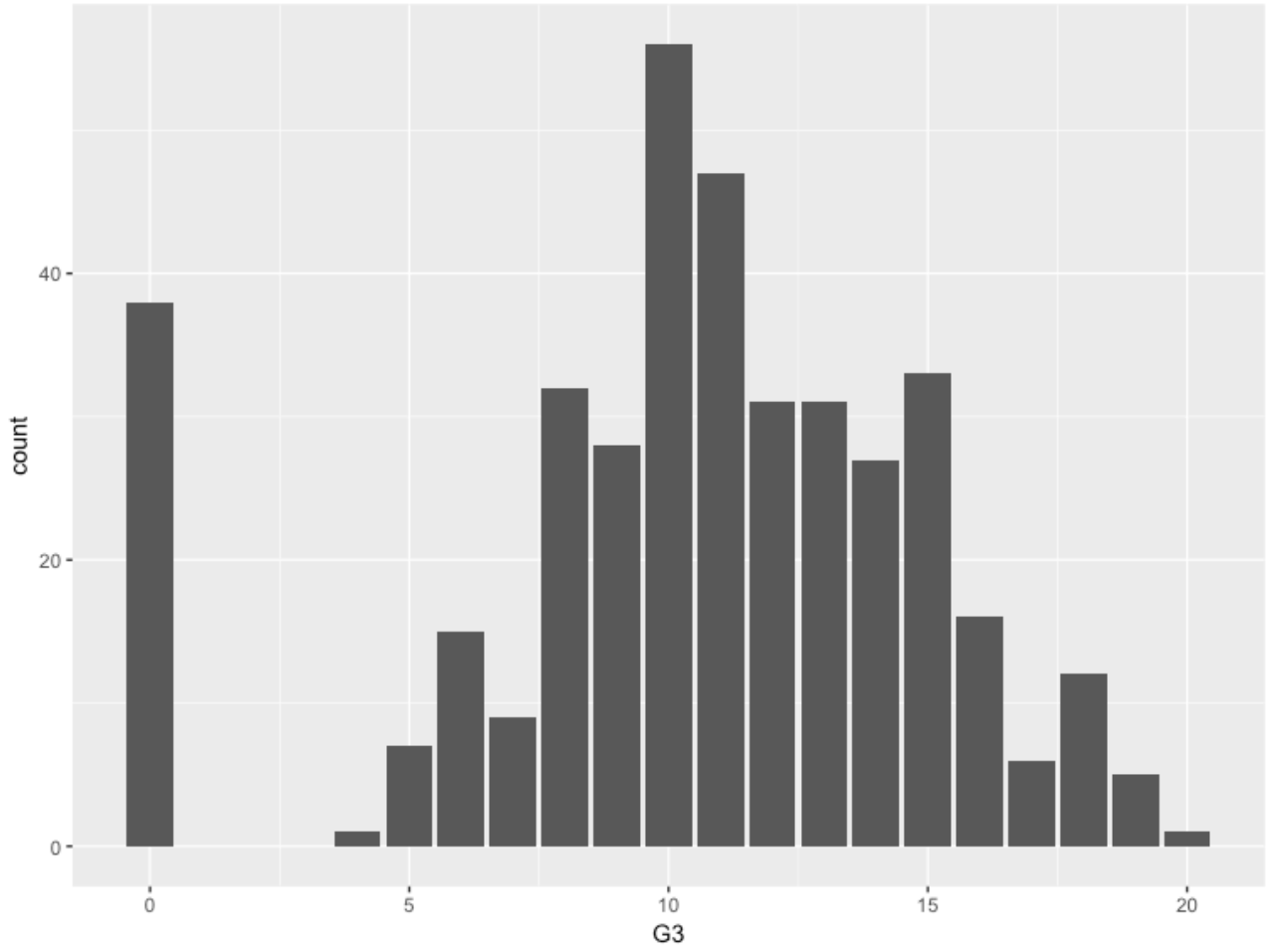
- Qualidade das relações familiares
- Quanto tempo livre possui fora da escola
- Frequência com que sai com os amigos
- Frequência com que bebe em dias de semana
- Frequencies com que bebe em fins de semana
- Saúde
- Número de faltas
- Nota no primeiro período
- Nota no segundo período
- Nota final

Existem dois arquivos no .zip fornecido pelo Kaggle. O primeiro possui a nota dos alunos somente de matemática e possui 395 alunos. O segundo possui a nota somente de portugues, com 649 alunos. Dentre os alunos da base, 382 deles foram entrevistados para ambos.

Amostra base de dados 1:

school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no	no	no	yes	yes	no	no	4	3	4	1	1	3	4	0	11	11
GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes	no	no	no	yes	yes	no	5	3	3	1	1	3	2	9	11	11
GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	0	yes	no	no	no	yes	yes	yes	no	4	3	2	2	3	3	6	12	13	12

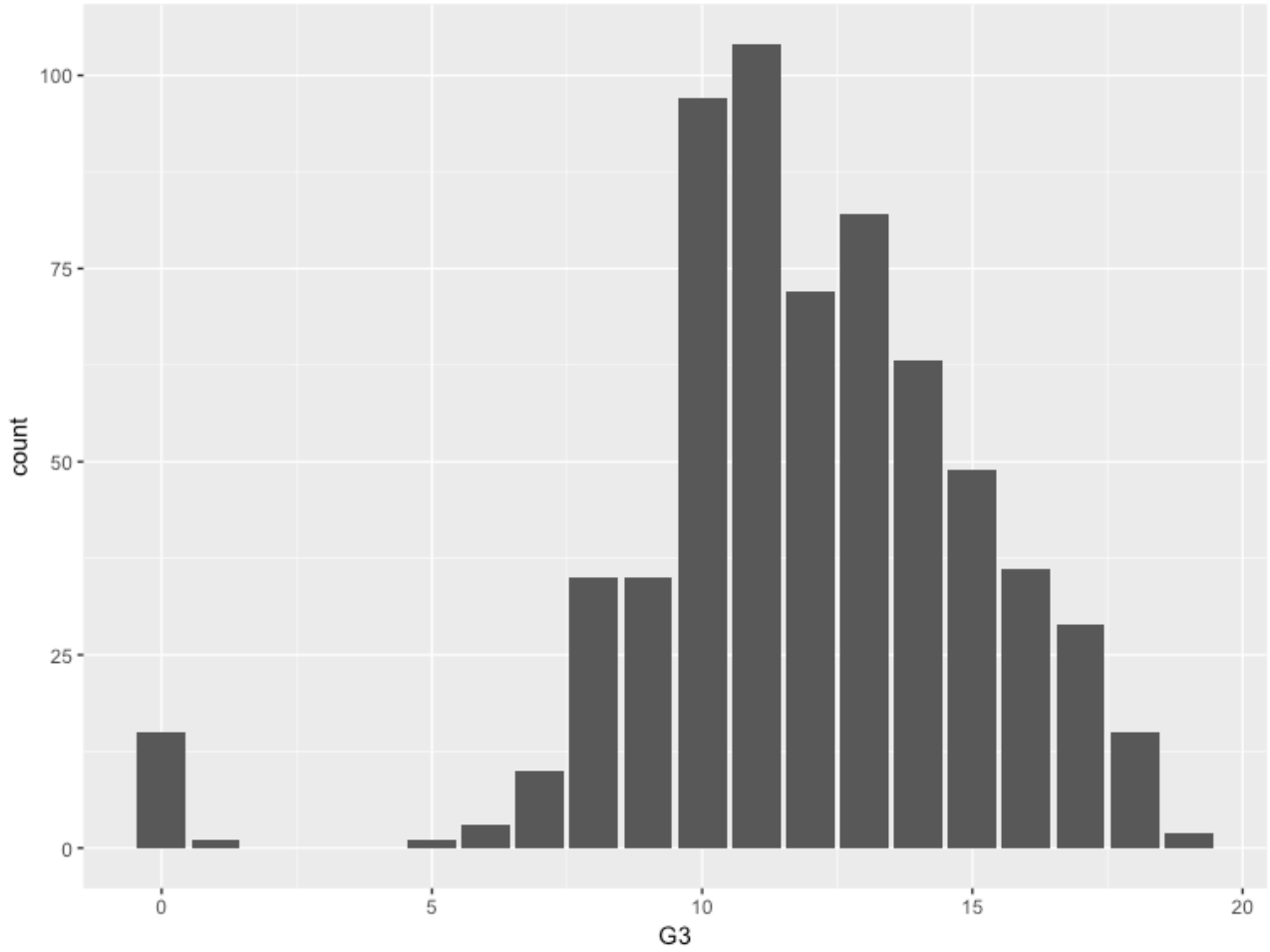
Distribuição das notas finais da base 1:



Amostra da base de dados 2:

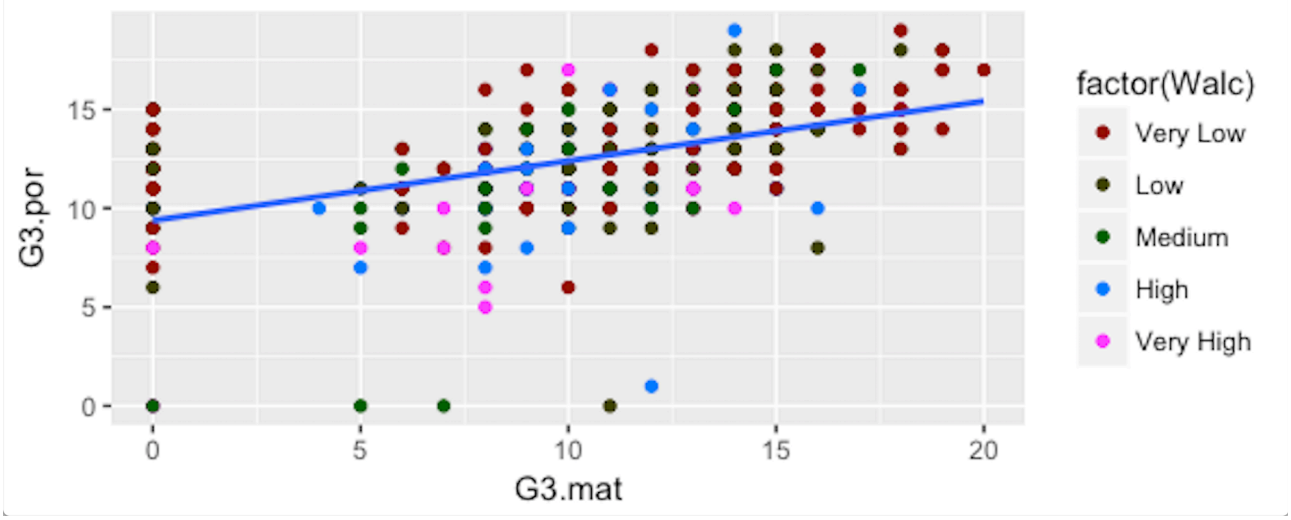
school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no	no	no	yes	yes	no	no	4	3	4	1	1	3	6	5	6	6
GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes	no	no	no	yes	yes	no	5	3	3	1	1	3	4	5	5	6
GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	3	yes	no	yes	no	yes	yes	yes	no	4	3	2	2	3	3	10	7	8	10

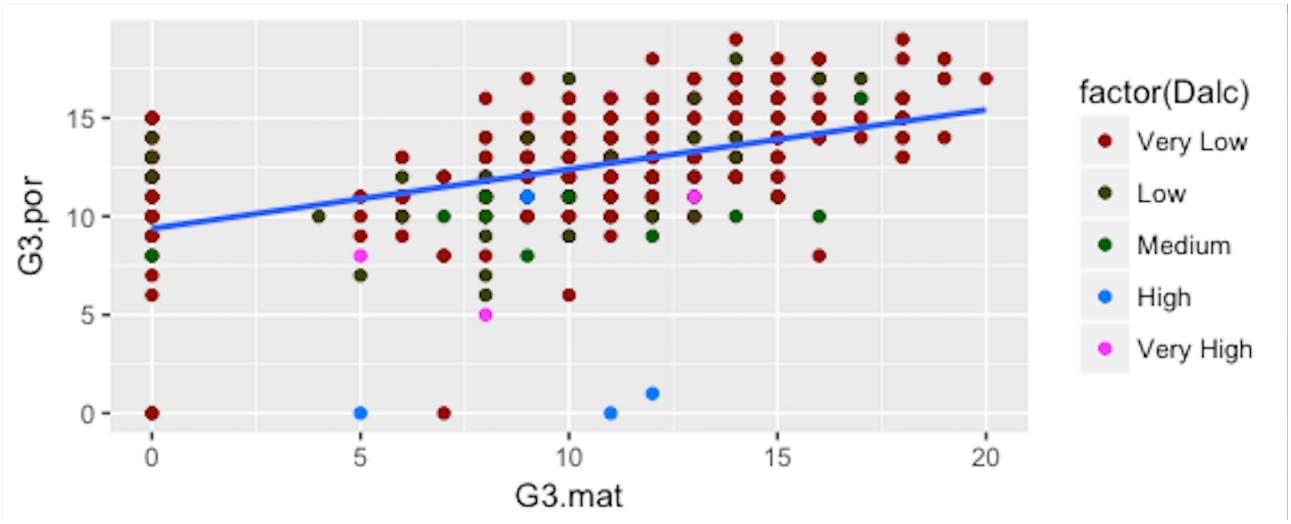
Distribuição das notas finais da base 2:



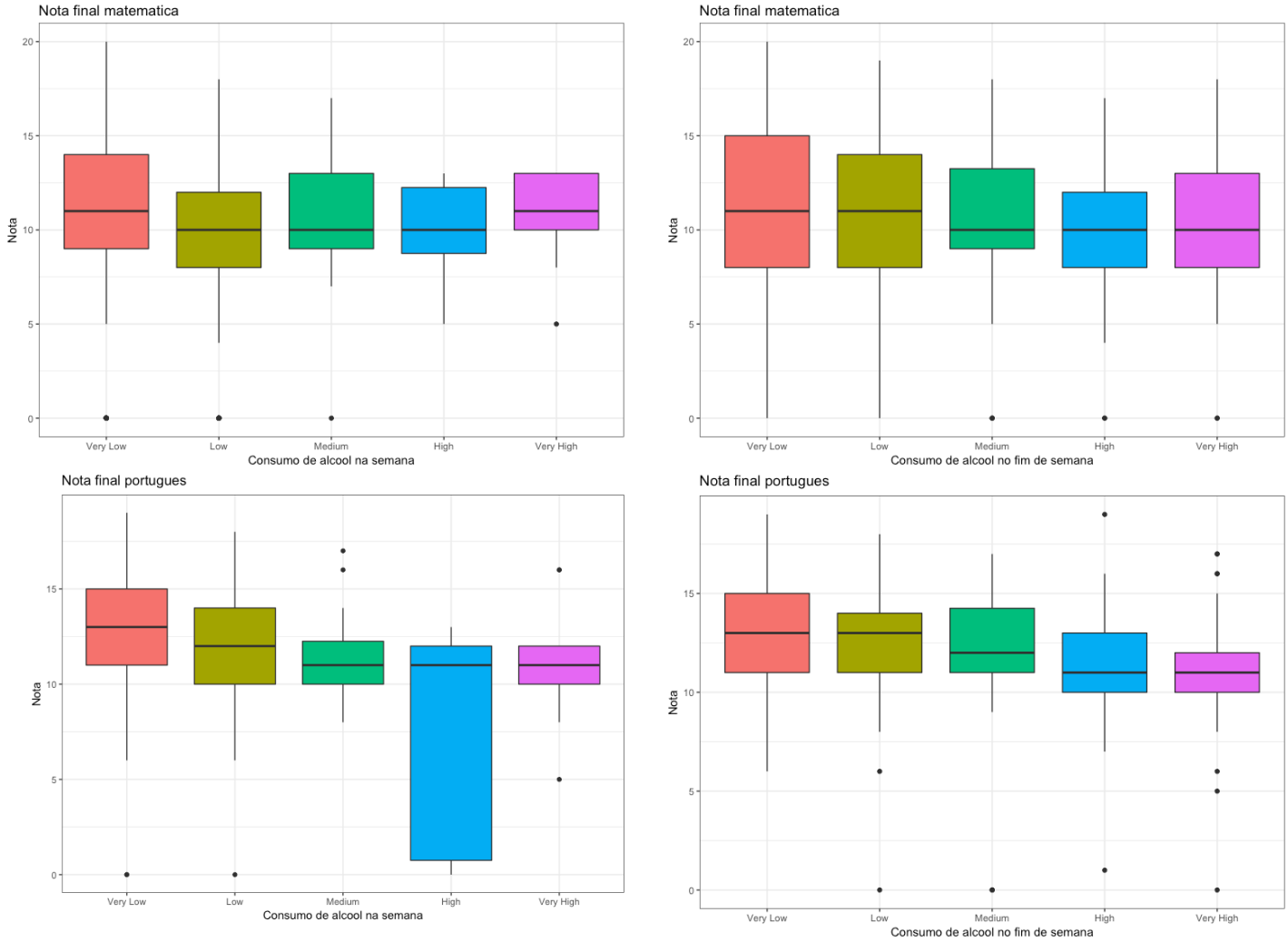
3. Metodologia:

No trabalho vamos primeiro agrupar as bases com notas de português e matemática em uma só. Após isso, vamos visualizar a distribuição dos estudantes





usando as notas finais de portugues, matemática e o consumo de álcool como parâmetros.
Lembrando que Walc é o consumo de álcool dos estudantes durante o fim de



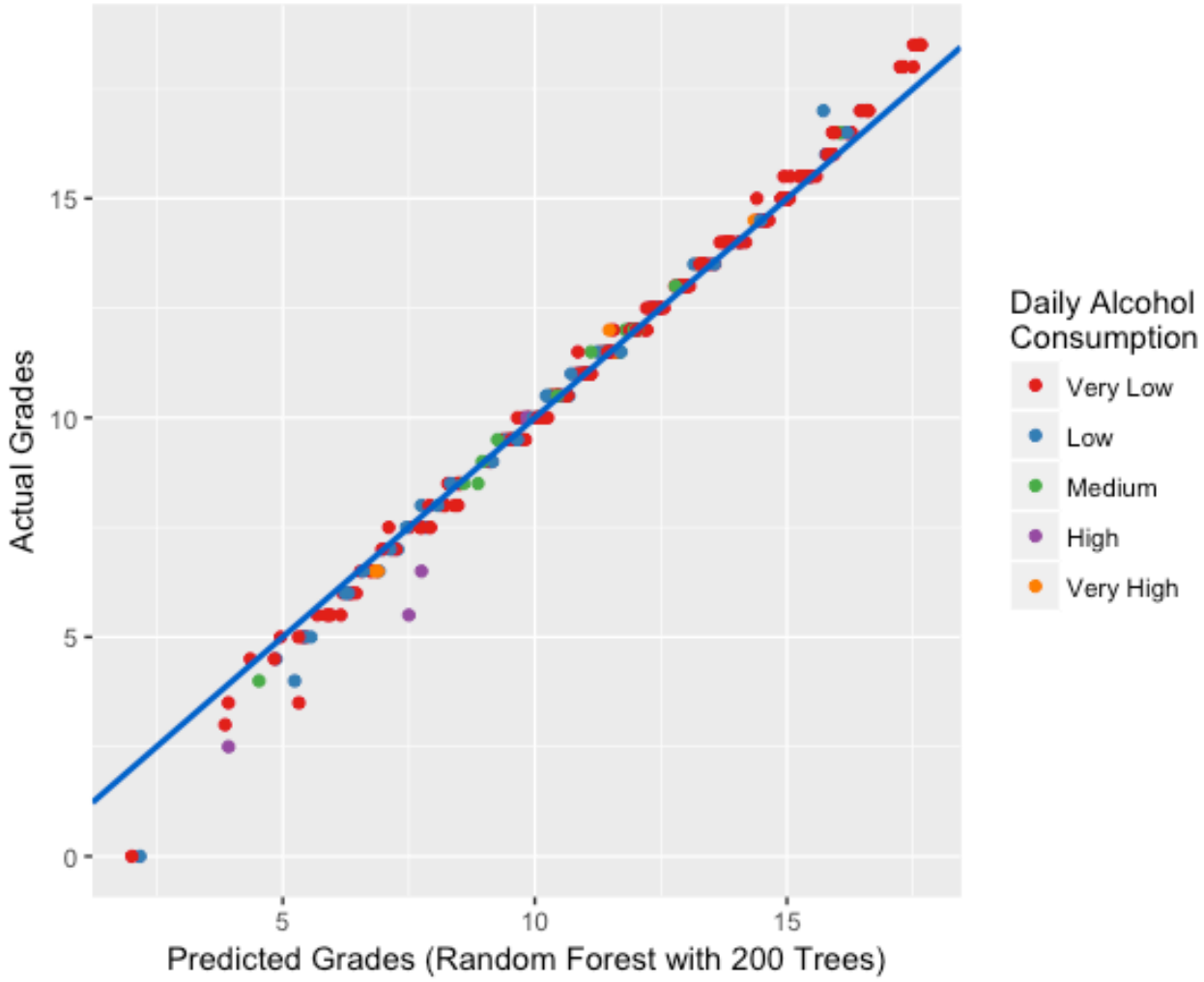
semana e Dalc é o consumo de álcool
durante a semana. Para visualizar melhor se há relação entre o consumo e as notas, vamos fazer outros quatro gráficos.

Percebe-se que há uma correlação forte entre um consumo de álcool alto durante a semana e notas baixas de português, contudo, em alunos com o consumo muito alto o mesmo não acontece. Isso sugere que seja coincidência a correlação.

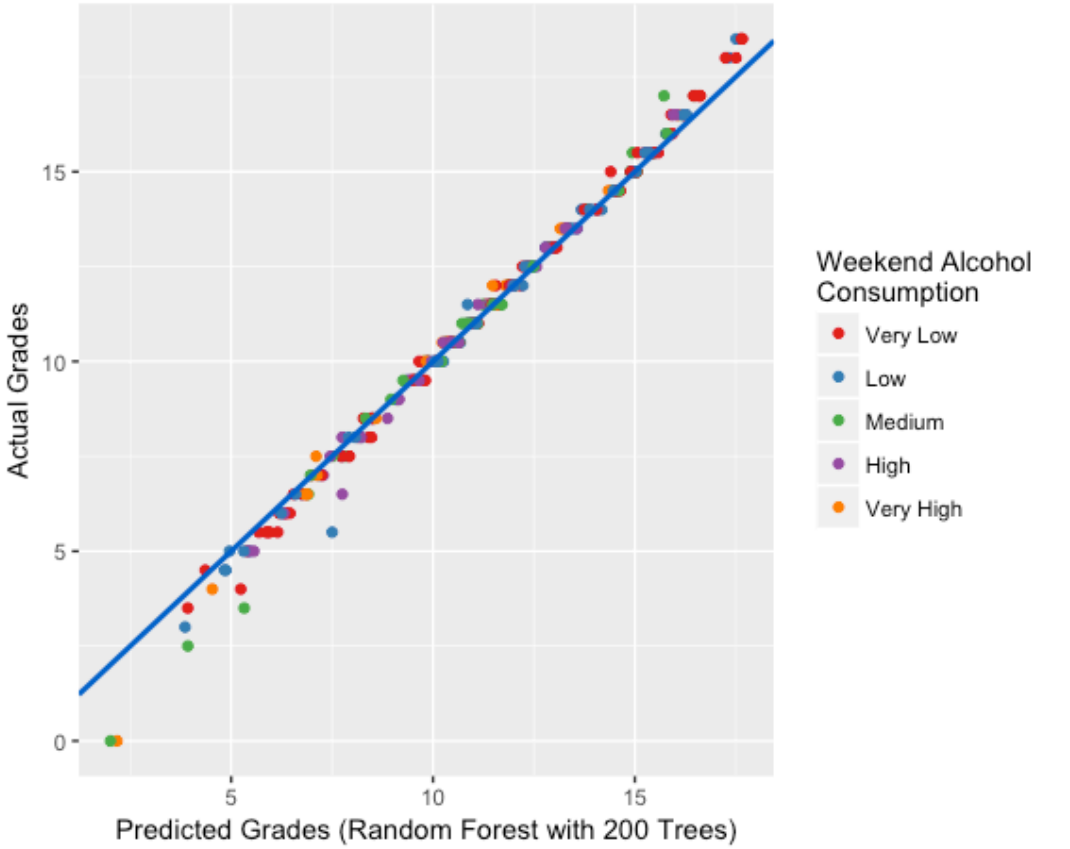
Agora que já visualizamos como estão distribuídos os alunos, vamos criar um modelo de classificação usando o algoritmo Random Forest e vamos analisar quais são as features mais importantes para a previsão de acordo com o resultado do algoritmo. Para o algoritmo de random forest, será passado como parâmetro de previsão a média entre as notas de português e matemática por simplicidade. Nesse trabalho a base não será dividida em bases de teste e de treino pois a intenção não é ver se o modelo consegue prever corretamente a nota dos alunos, e sim entender se há relação entre o consumo de álcool e as notas dos alunos e, se houver, saber qual a relevância dessa relação.

4. Avaliação:

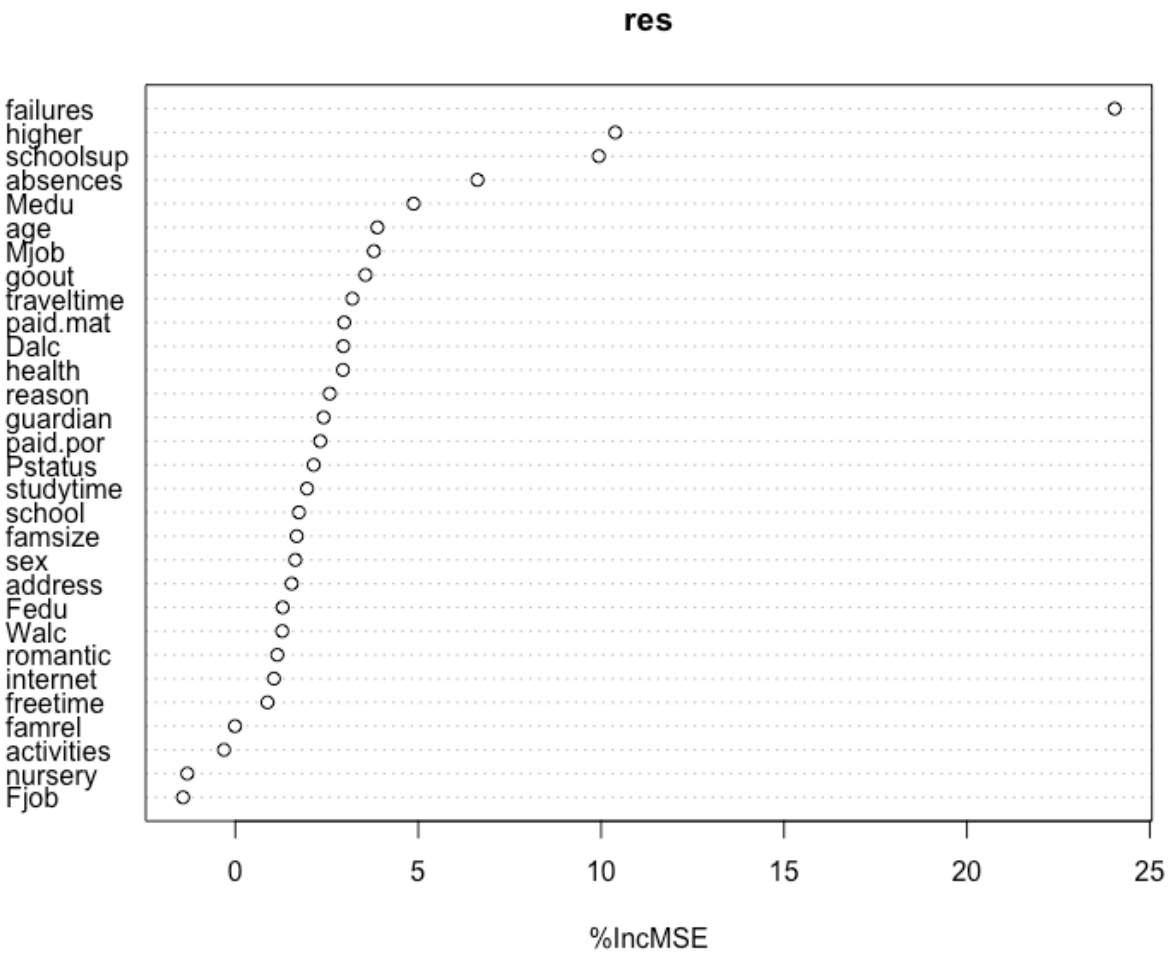
Ao ser rodado o random forest com 200 árvores, o modelo obteve sucesso em prever as notas.



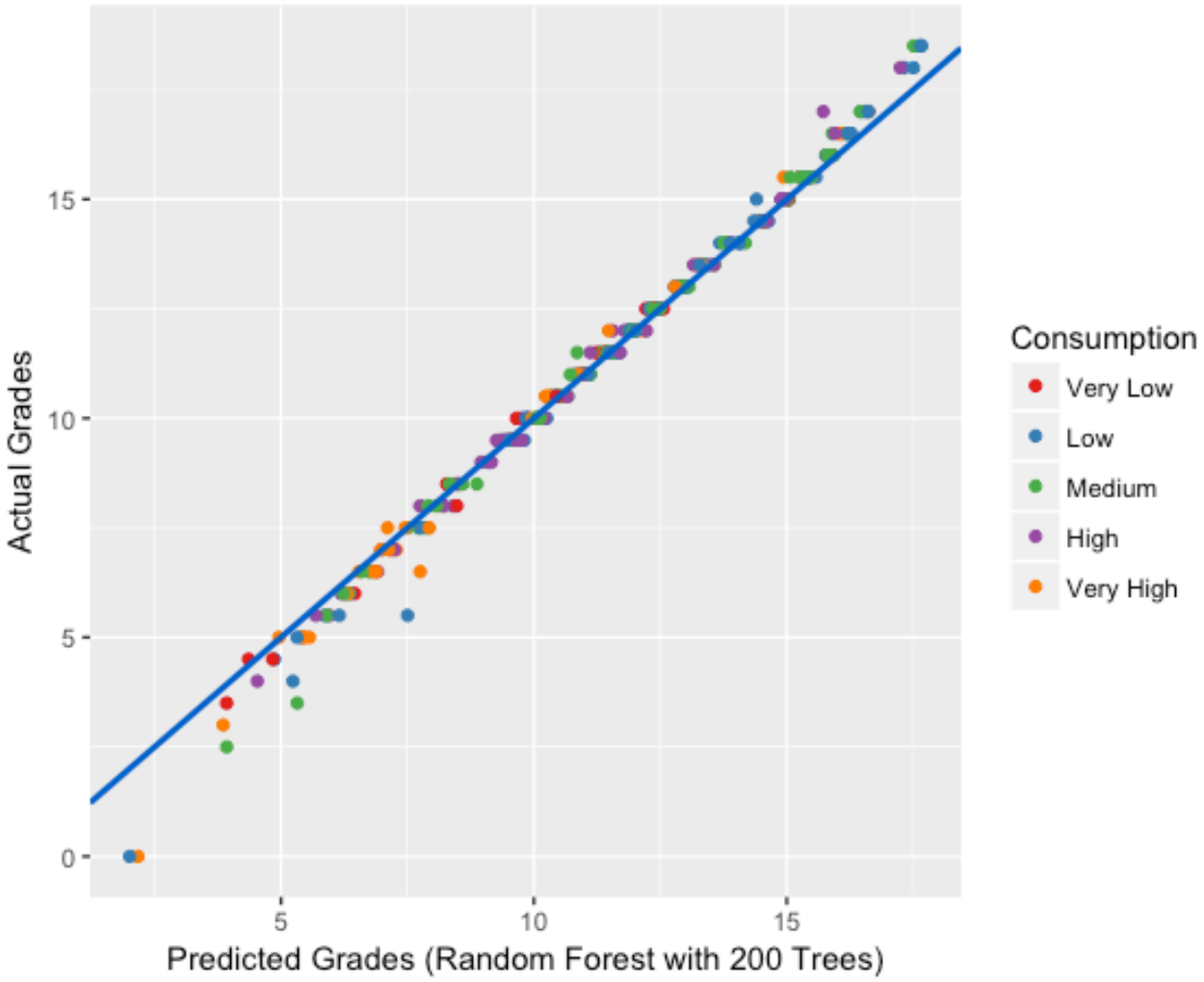
Parecem ter menos alunos com consumo diário de álcool baixo com notas altas do que alunos com consumo diários de álcool alto com as mesmas notas, deve haver uma correlação. Vamos testar se a mesma correlação existe no caso de consumo de álcool nos fins de semana



A correlação se torna menos aparente no caso do consumo de álcool nos fins de semana, o que faz sentido, uma vez que pessoas que estão bêbadas durante a semana tem menos chance de estar estudando do que pessoas que estão sóbrias. Vamos analisar então quais foram as features mais relevantes para o algoritmo



O fato do número de repetências na matéria ser a variável mais importante é surpreendente, mas é simples de se explicar, pois alunos que tomam bomba geralmente são menos esforçados. Features como higher, schoolsup, absences dispensam explicações. A educação e trabalho da mãe é mais relevante pois dos 382 alunos, 275 deles tem a mãe como guarda, enquanto apenas 91 deles tem guarda do pai. Provavelmente seria mais interessante levar em consideração apenas a educação e trabalho do guardião ao invés de levar em consideração do pai e da mãe. O tanto que os alunos saem com os amigos também parece influenciar as notas dos alunos, convivência social parece ter uma relação muito significante com as notas dos alunos, é possível perceber isso no gráfico abaixo:



Aulas particulares de matemática também parecem ajudar os alunos com as notas, ainda que as de português pareçam ser menos significantes. Isso provavelmente acontece porque 177 alunos tomam aulas particulares de matemática enquanto apenas 26 as tem de português.
Percebemos então que o consumo de álcool durante a semana possui apenas 2.95% de relevância para o modelo, está no top11 de features mais relevantes, ou seja, possui relevância significativa mas a mudança nela não faria uma diminuição tão significante. Já o consumo de álcool nos fins de semana se torna muito menos relevante para o modelo, com apenas 1.28% de importância. Vamos criar um modelo sem as variáveis relacionadas ao consumo de álcool para testar se os

valores de RMSE, R-Squared e Mean Absolute Error vão se alterar significantemente.

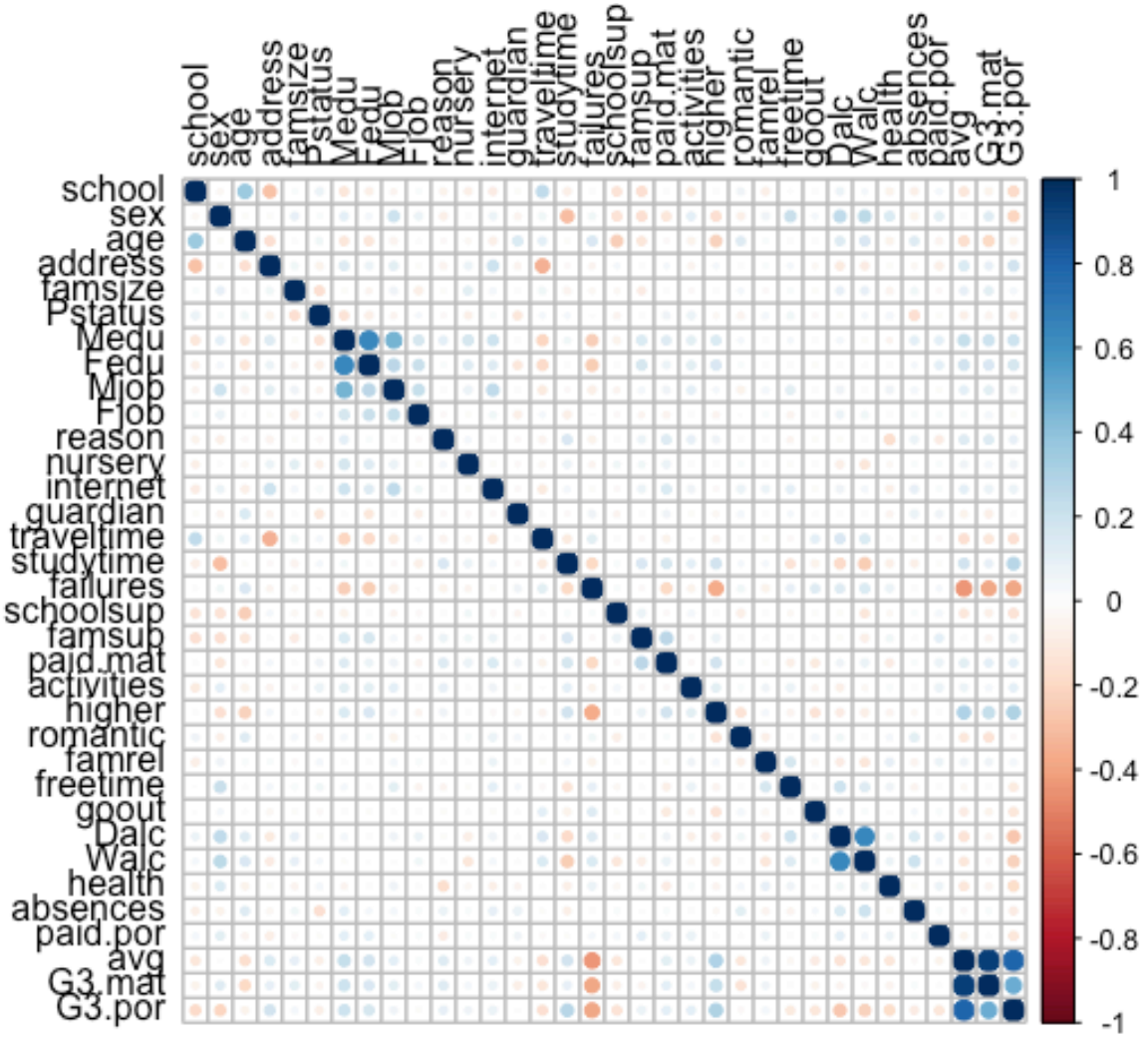
Modelo com variáveis relacionadas ao consumo de álcool:
RMSE: 2.6859
R-Squared: 0.3445
MAE: 2.0606

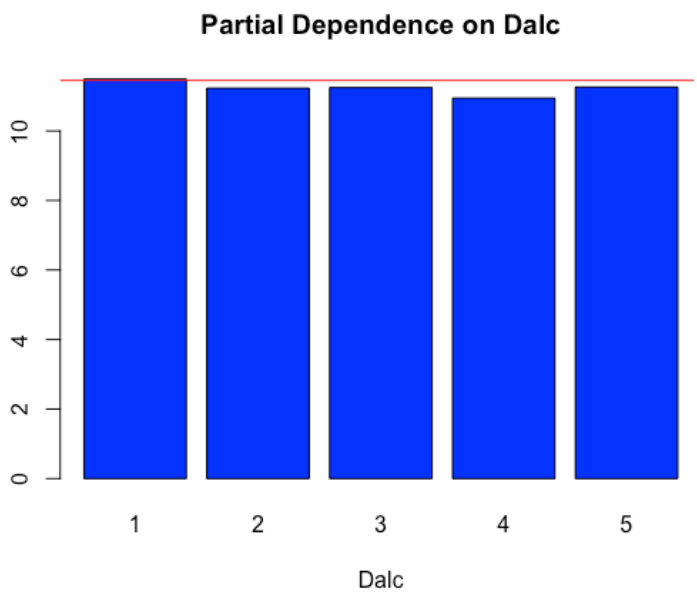
Modelo sem as variáveis de consumo de álcool:
RMSE: 2.7043
R-Squared: 0.3355
MAE: 2.0743

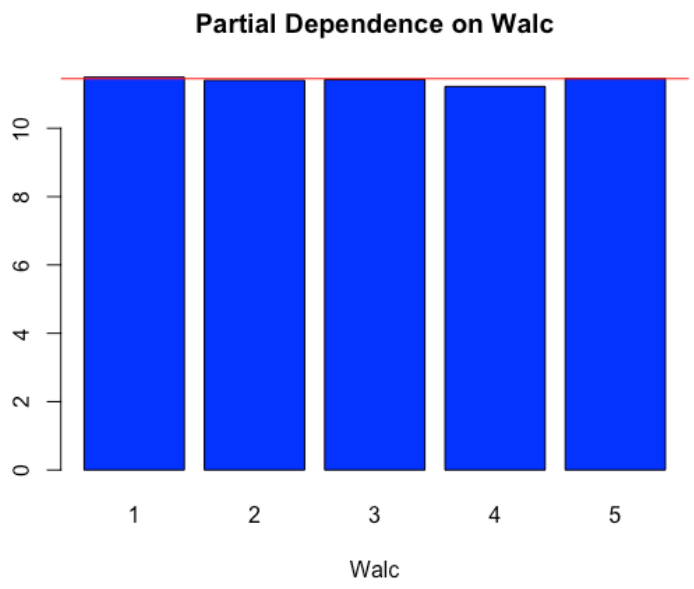
Percebemos que de fato a diferença entre ambos é pequena.

5. Conclusão:

É possível perceber que o consumo de álcool não possui uma correlação muito significante com notas dos alunos. Percebe-se isso com a matriz de correlação a seguir







Percebemos com todo esse trabalho que a relação do consumo de álcool com o desempenho escolar não é tão significante, especialmente se for um consumo baixo durante a semana.

6. Referências:

- https://www.kaggle.com/uciml/student-alcohol-consumption
- http://www.cookbook-r.com/Graphs/
- https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html
- https://www.r-bloggers.com/random-forests-in-r/
- https://stats.stackexchange.com/questions/189906/accuracy-of-training-sample-in-random-forest-model-in-r
-