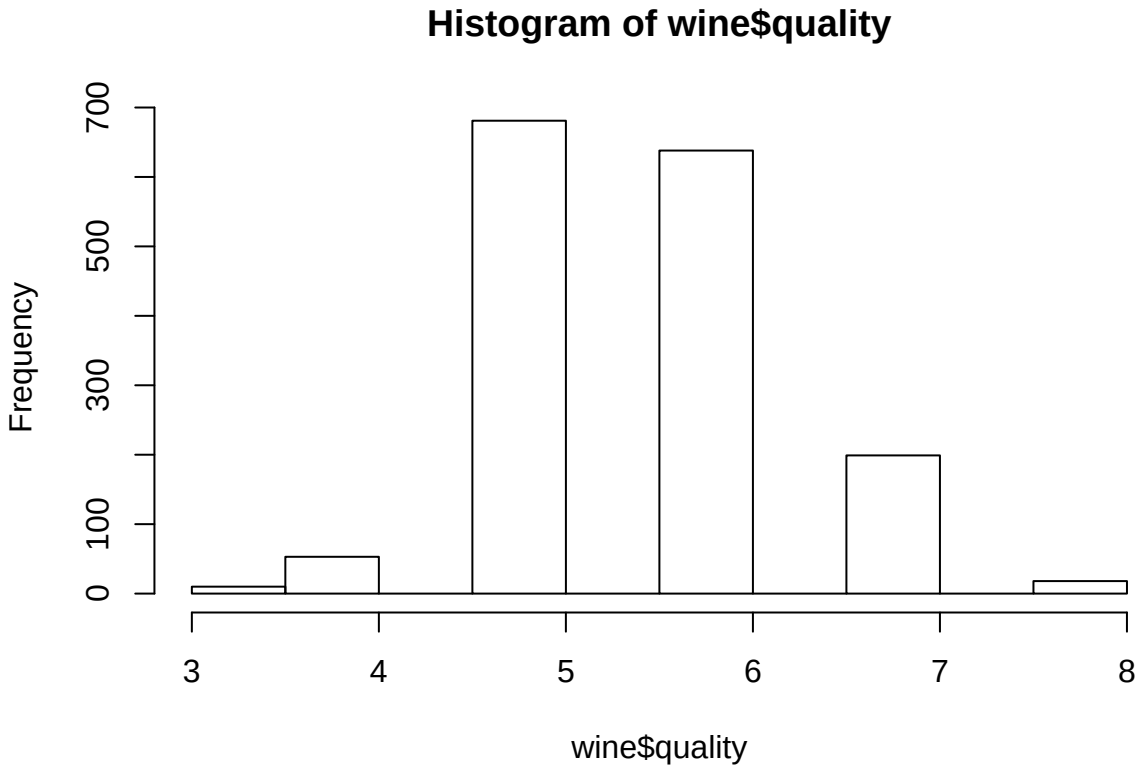
Homework 3

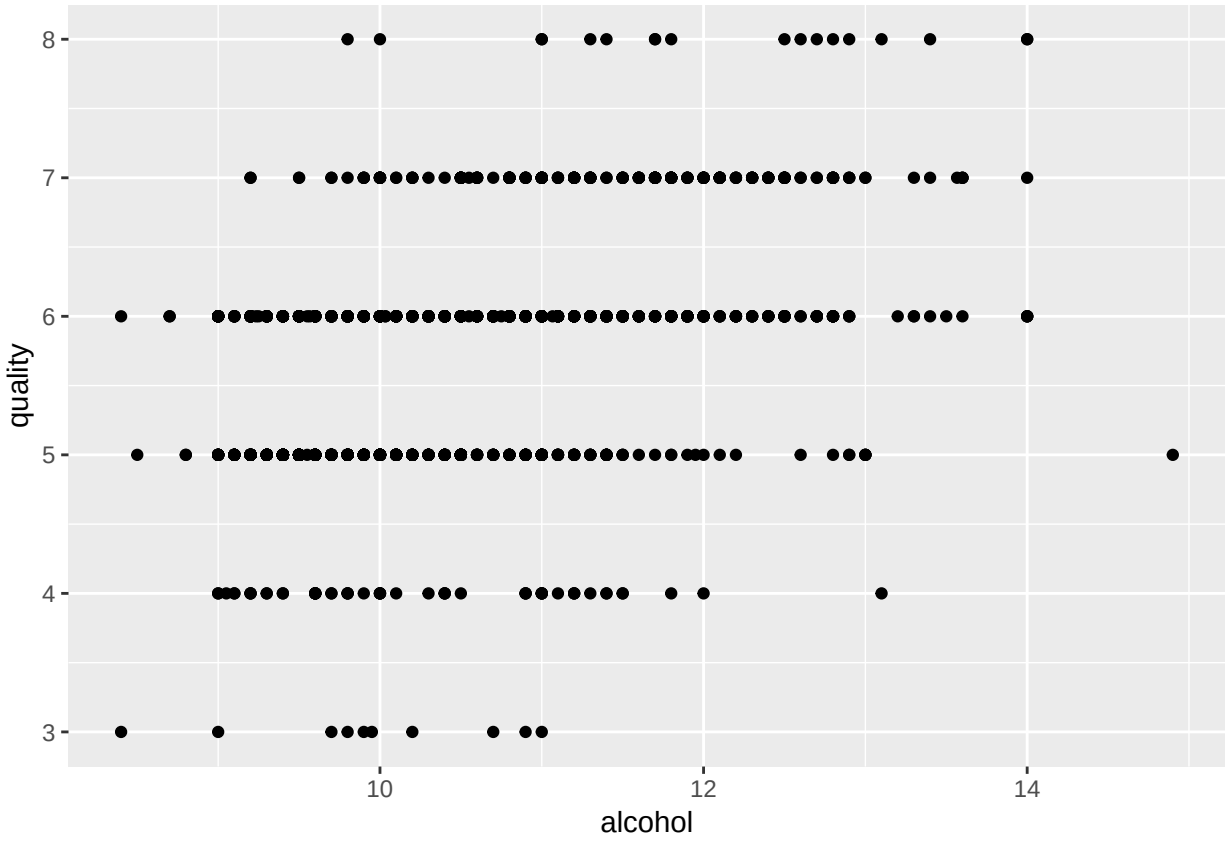
Introduction

You now know everything that's necessary to do some predictive modeling. In this exercise, we will go through the process again and reinforce our knowledge of regression.

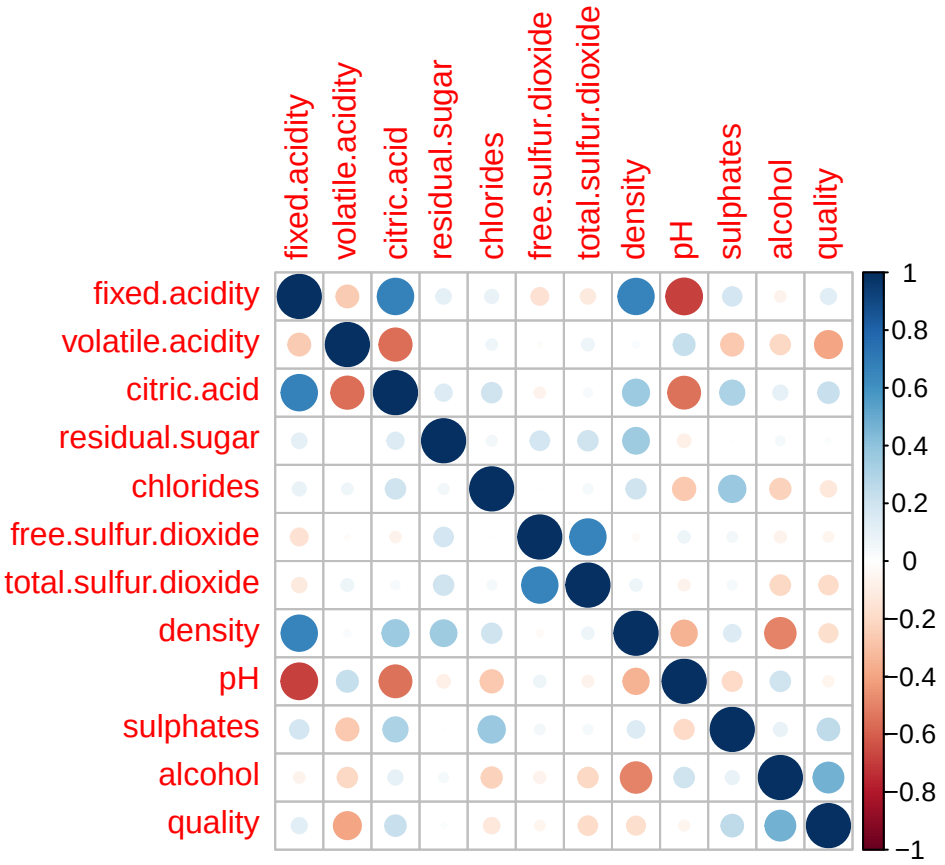
1. We will use a dataset from UCI repository https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv, that provides information about the red wine characteristics. The target variable is quality (score between 0 and 10). The variable names are fairly self explanatory.
2. Look into the data, how many observations does your dataset have? Any missing values? Do you think we should keep the target variable numeric?
3. Get a feeling for the data with some easy visualizations (use ggplot2). Start with some scatterplots, see how the quality score is distributed.
4. Given that all variables are numeric, getting an idea about the internal correlation structure would be a good step. Which factors seem to have most influence?




Histogram of wine$quality
Frequency
0
100
300
500
700
3
4
5
6
7
8
wine$quality




```
## corrplot 0.84 loaded
```



5. Let's try to predict the quality of wine, using the data about it's chemical compounds. Build a simple linear regression, start only with two components that seemed to have most influence (use **target_var ~var1+var2** formula). Look into the results. How well does your model explain the result?

6. Now let's try to use all our variables. Did the result get better? Which variables seemed to play major role? Did it match the guess we made after looking at correlation? Why could that be?