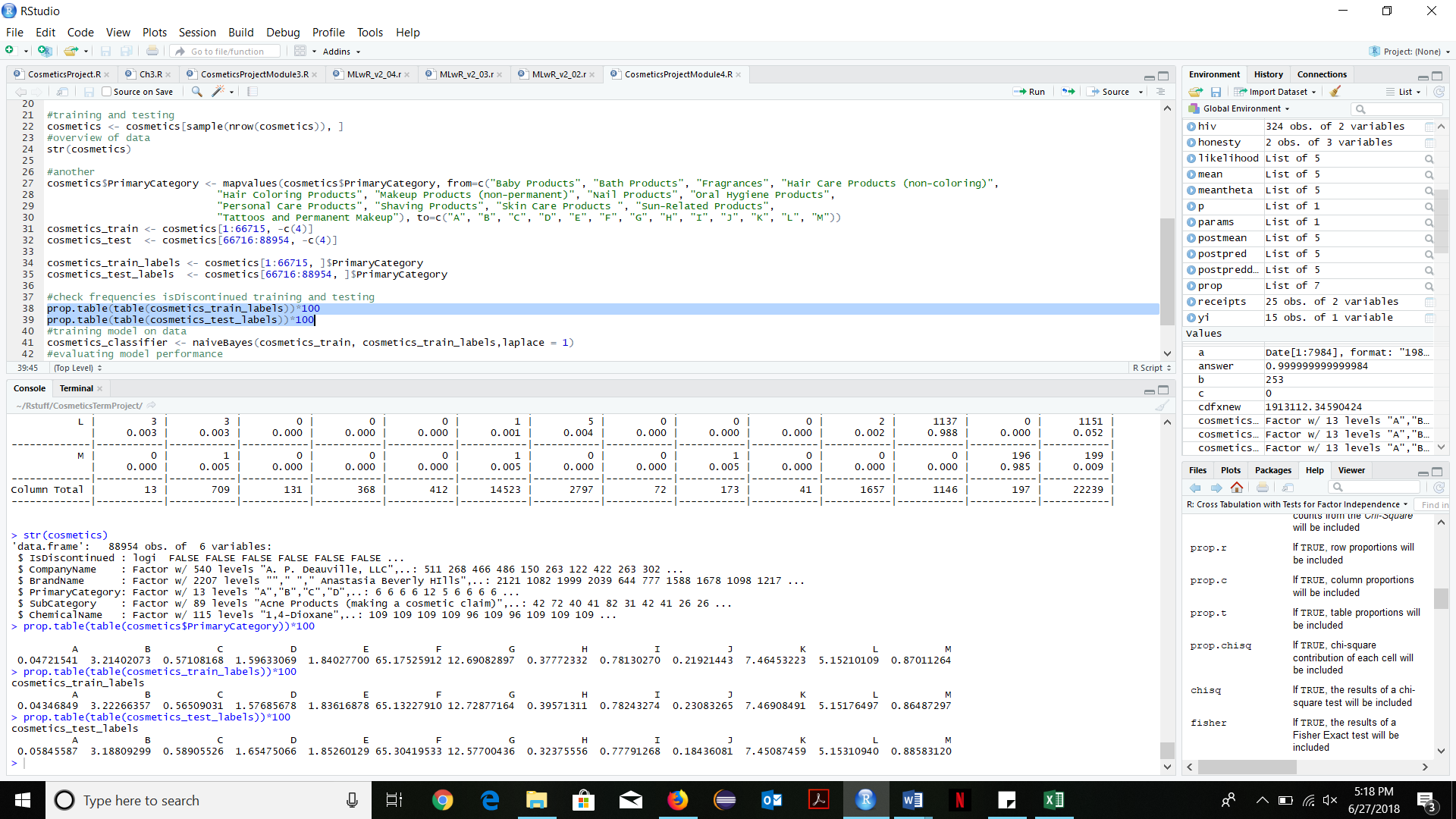
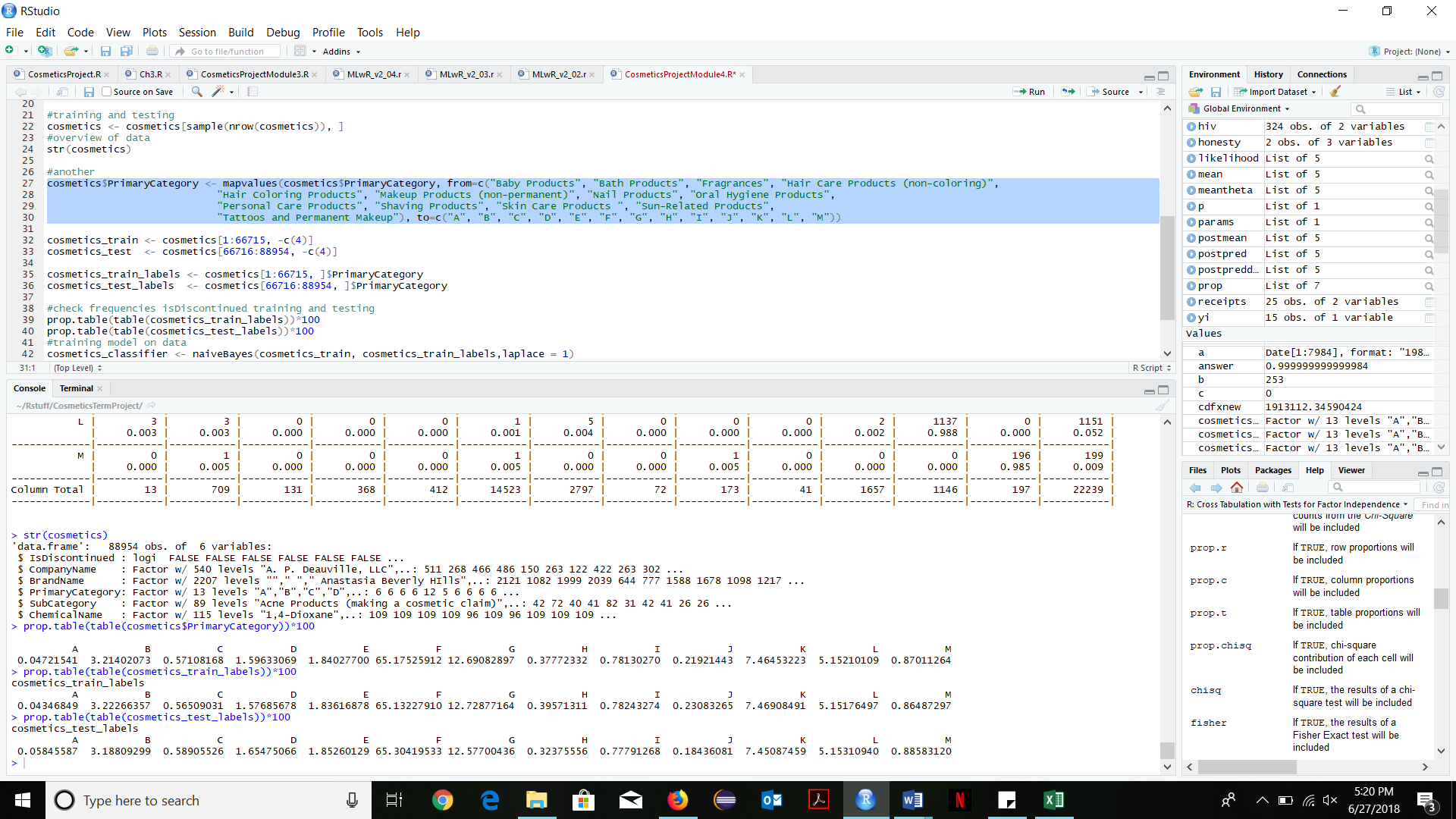
Module 4

I predicted the PrimaryCategory type of the cosmetic product based on the five categorical attributes: Company Name, Brand Name, SubCatergory, Chemical name, and isDiscontinued (Boolean). After randomizing the data (`cosmetics <- cosmetics[sample(nrow(cosmetics)), ]`), I used 75% (1:66715) of the data for training and 25% (66716:88954) for testing. As can be seen in the proprtion tables, the proportion of categories in the original data, the training data, and the testing data is about the same.



After seperating the training and testing data, I ran the Naïve Bayes algorithm on it setting the laplace estimator to 1 (I also tried with laplace=0 but had better results with laplace=1), and my results are in the crosstable. There were 13 different types of PrimaryCategory products. I changed the names to A-M since the names were too long to fit into a crosstable that could be viewed in one page.



The diagonal in the crosstable shows the proportions of products that were predicted to be in the right category and in all of the categories, more than or equal to 90% (rounded off to the nearest digit) of them have been identified to be of that category. The only exception is category A (Baby Products) which has only 3 products and 2 of them have been identified correctly. All in all 22085/22239 of the results have been predicted correctly which is a whopping 99.3 percent of the test data, which is genuinely great and surprising!

