A Logistic Regression Approach to CoIL Challenge 2000

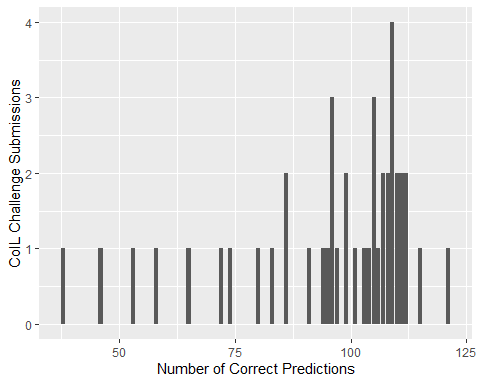
Corey Arnouts, Adam Douglas, Jason Givens-Doyle, Michael Silva

A logistic regression based solution to the CoIL Challenge 2000 is described. The challenge consists of correctly identifying potential customers for an insurance product, and describing their characteristics.

# Introduction

Businesses use data science to extract insights from data. It has many practical business applications. Identifying households to include in a marketing campaign is one application. One example using real world data is the Computational Intelligence and Learning (CoIL) Challenge. The CoIL Challenge competition was held from March 17 to May 8 in 2000. The challenge is to:

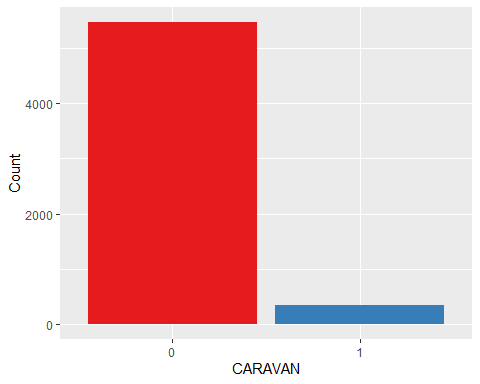
1. Identify potential customers for an insurance policy; and
2. Provide a description of this customer base.



In total 147 participants registered and 43 submitted solutions (Putten, Ruiter, and Someren 2000). The maximum number of policyowners that could be found was 238. The submissions identified 95 policy owners on average. The winning model (Elkan 2000) identified 121 policy owners. Random selection results in identifying 42 policy owners. The standard benchmark tests result in 94 (k-nearest neighbor), 102 (naïve bayes), 105 (neural networks) and 118 (linear) policy owners. (Putten, Ruiter, and Someren 2000). In this paper we set out to complete the first part of the COIL Challenge.

# Exploratory Data Analysis

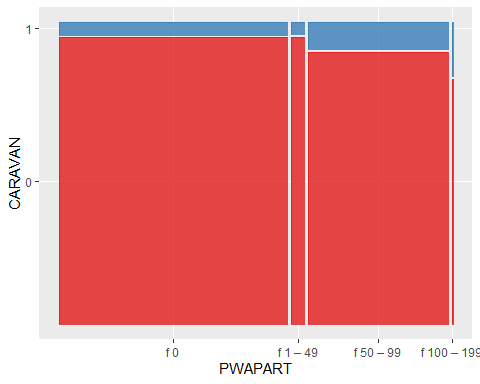
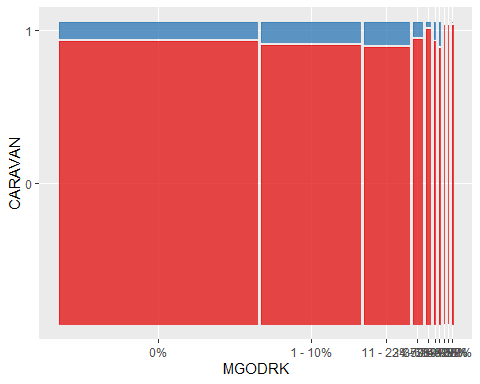
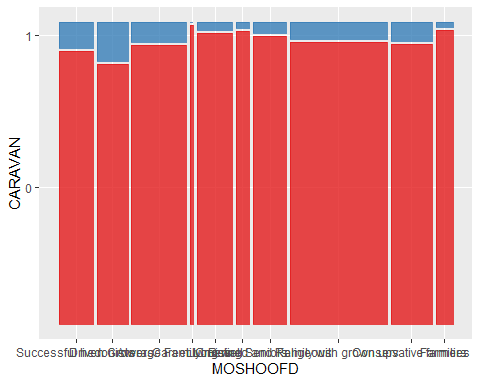
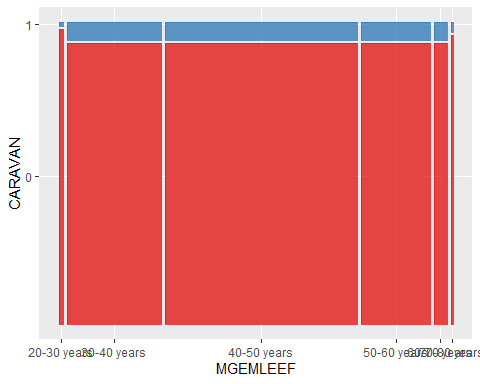
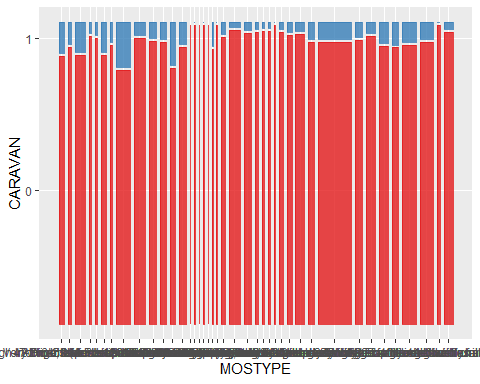
## Response Variable CARAVAN



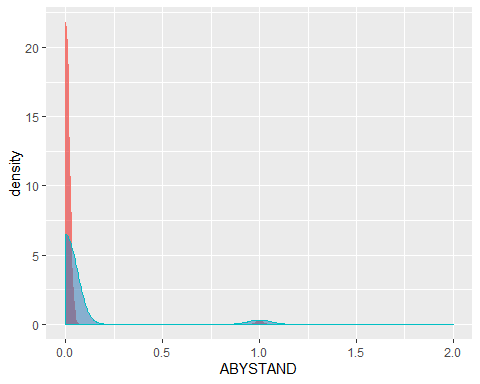
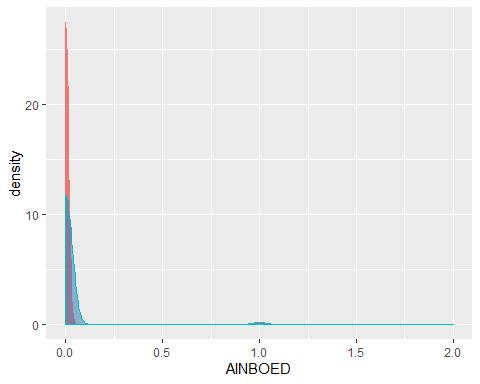
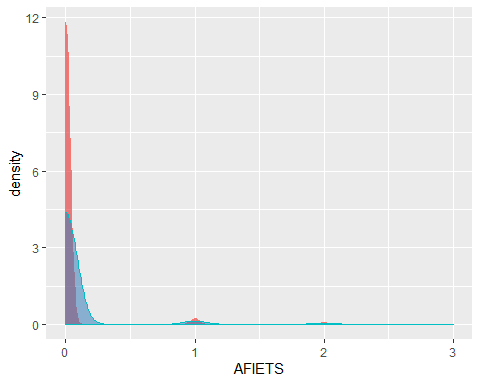
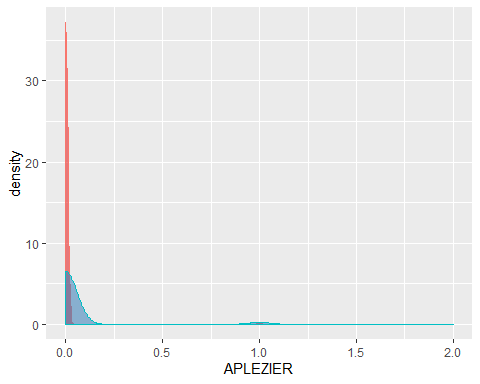
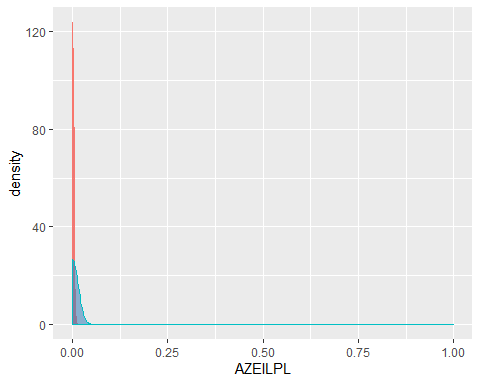
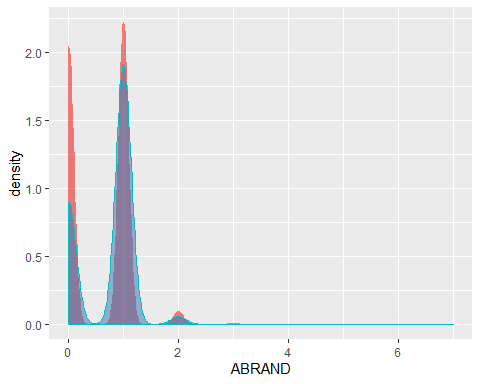
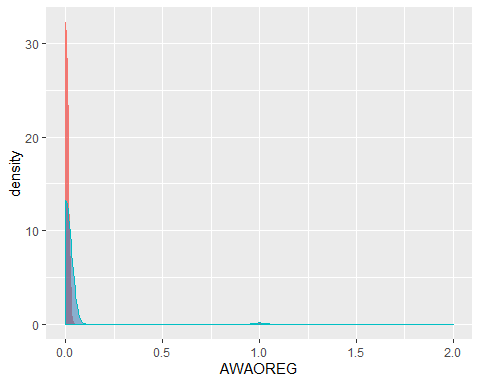
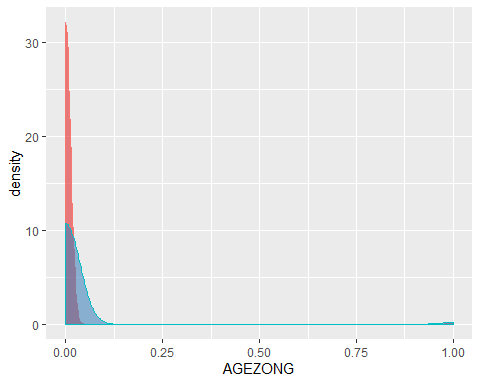
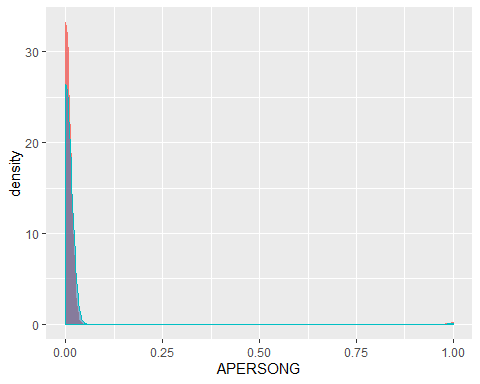
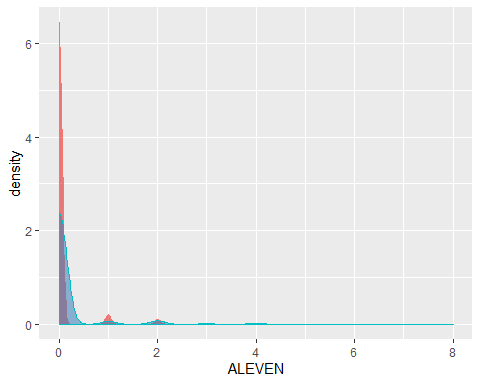
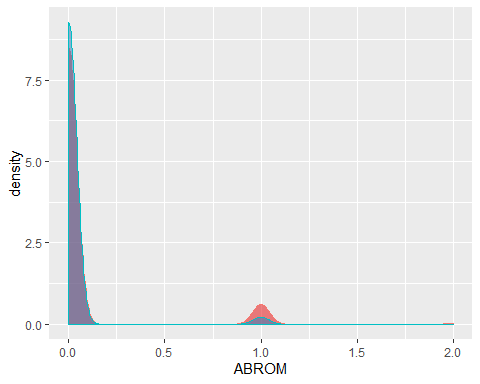
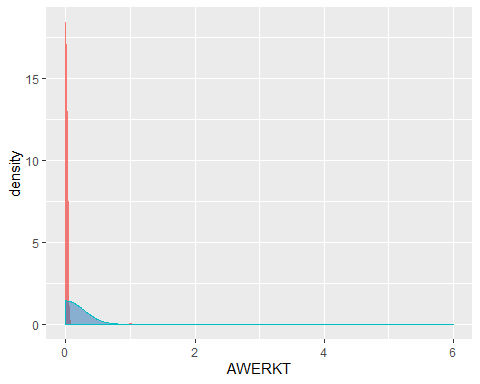
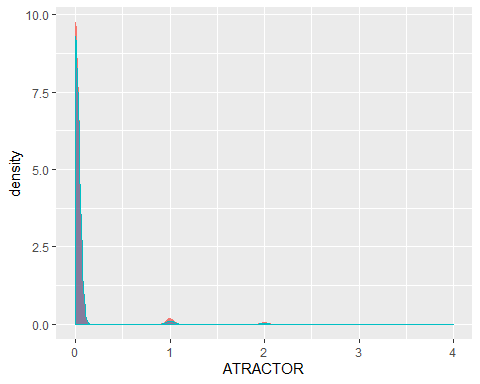
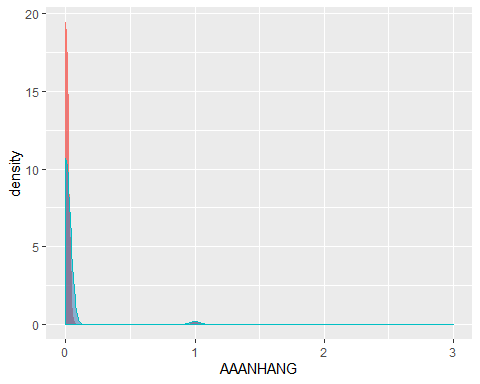
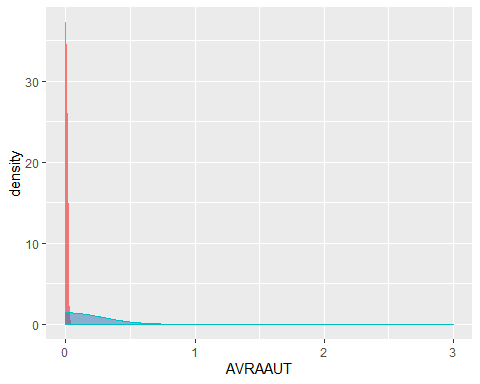
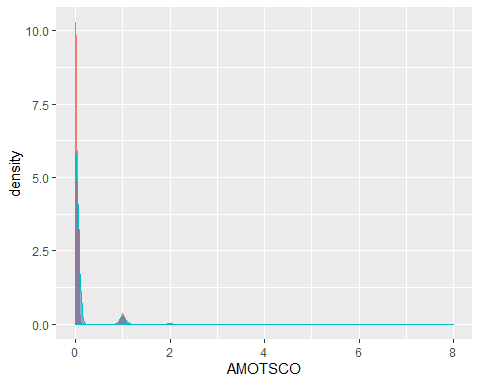
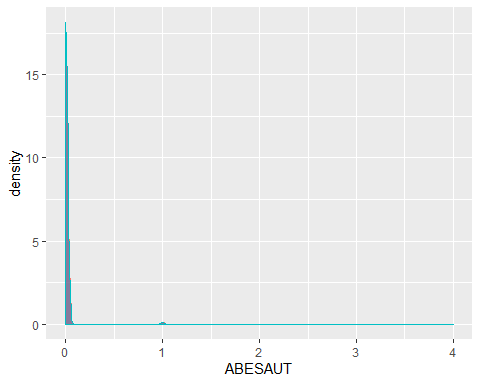
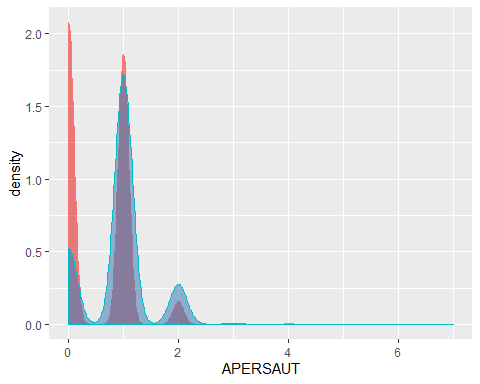
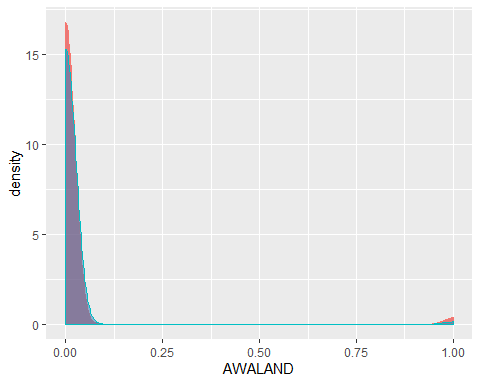
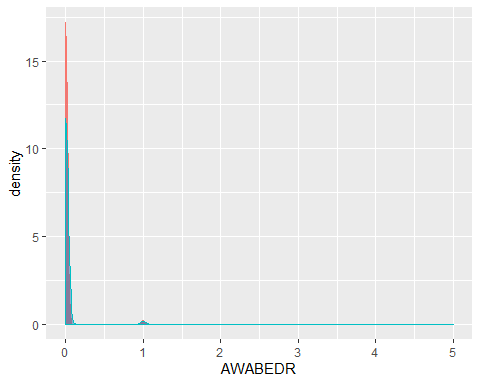
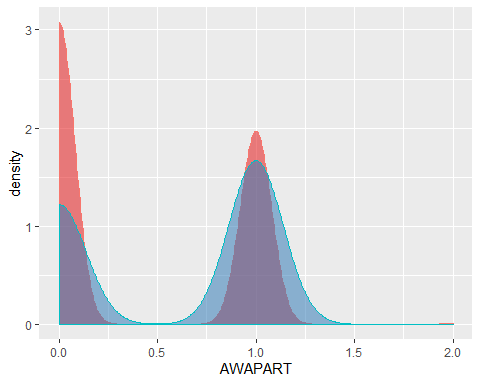
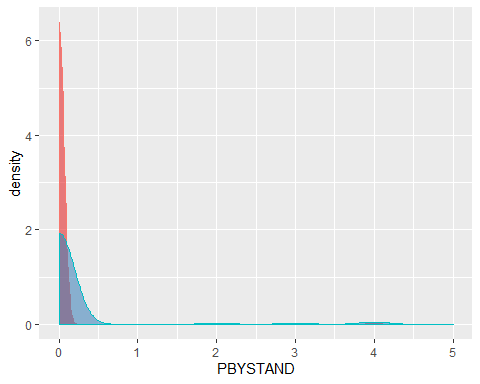
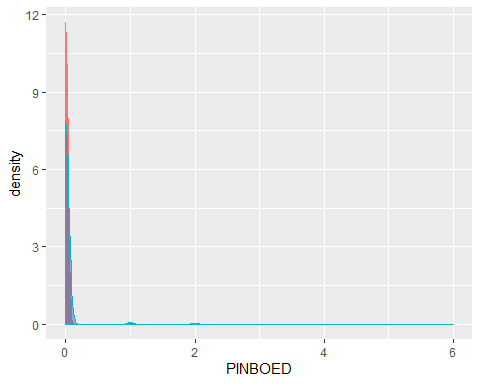
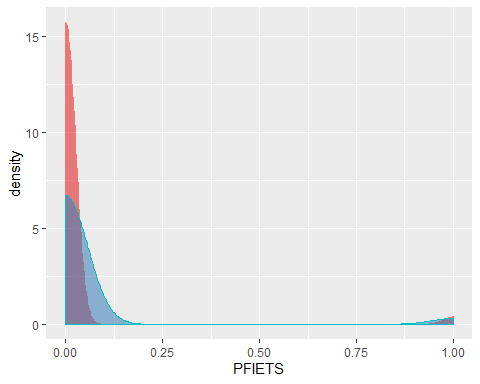
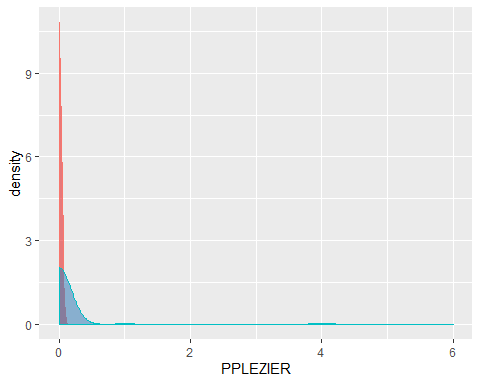
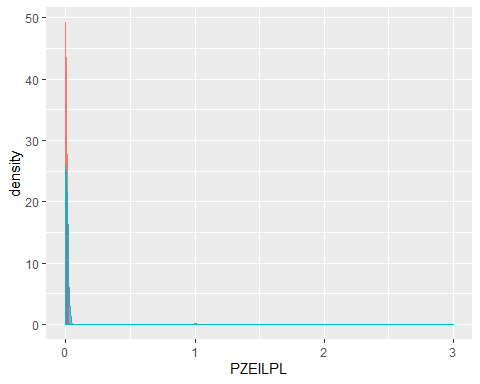
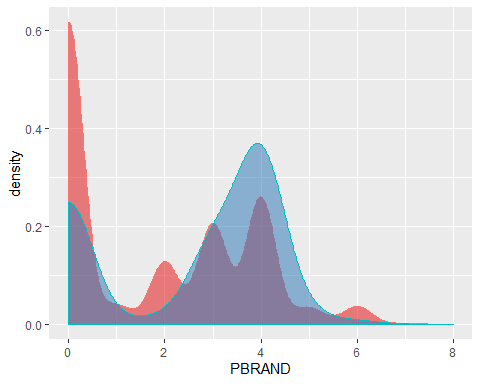
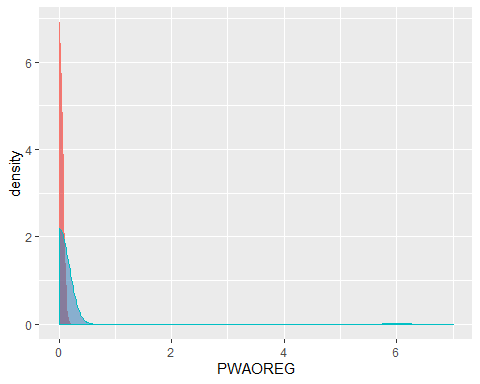
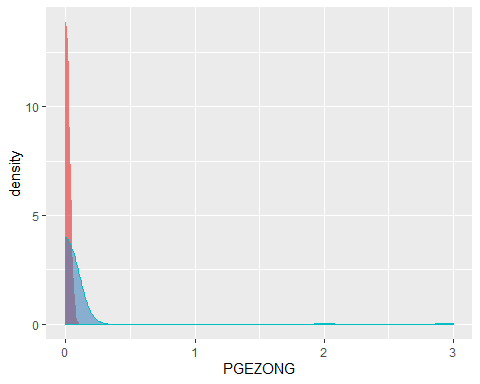
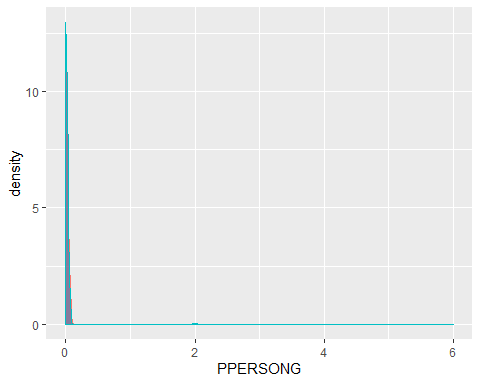
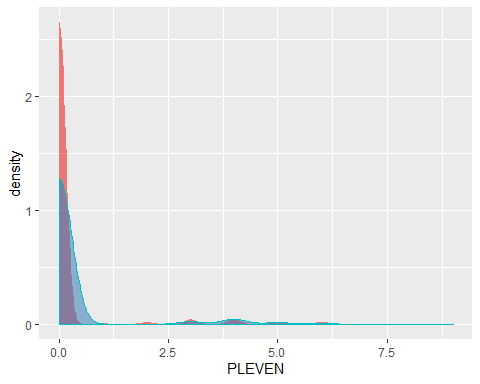
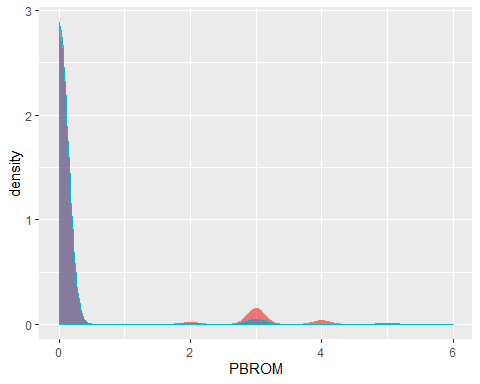
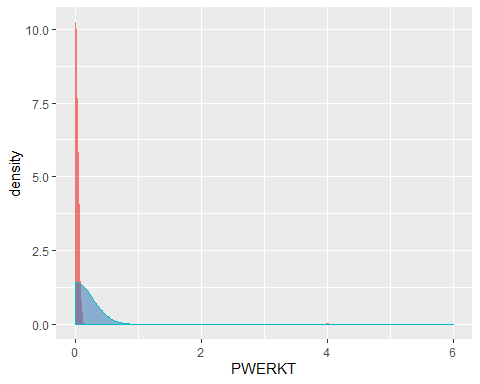
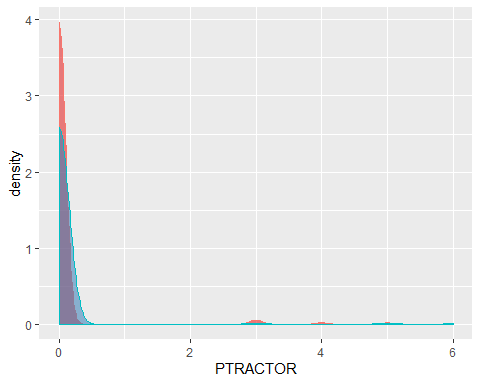
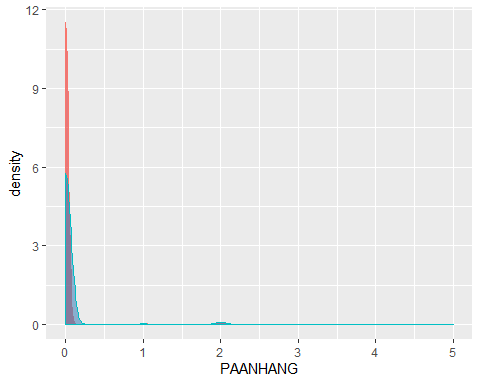
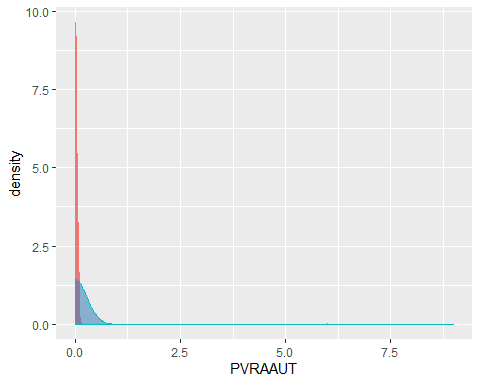
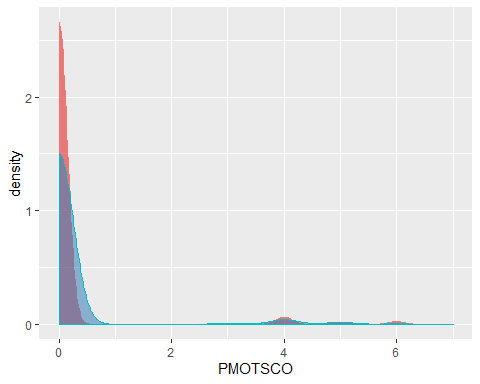
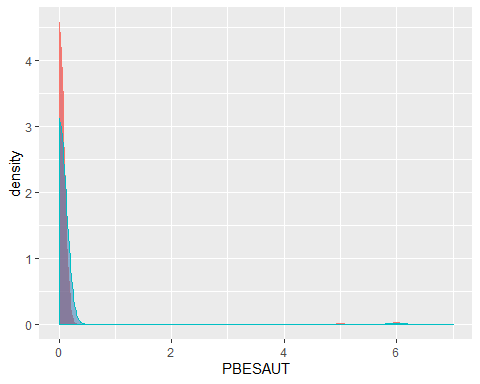
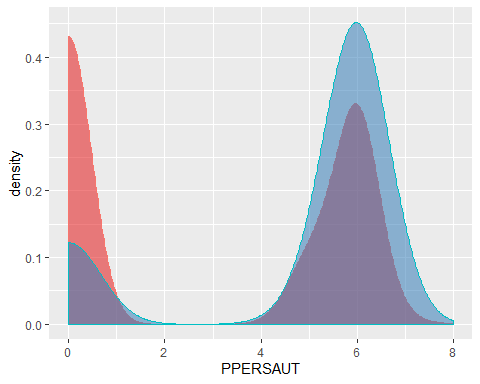
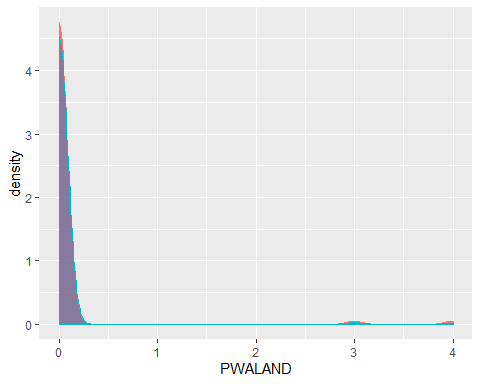
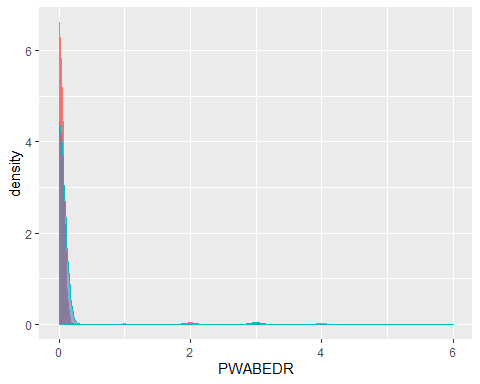
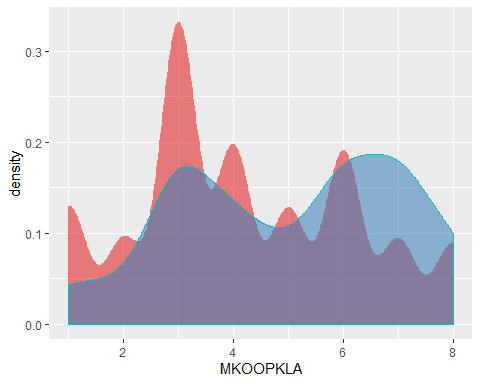
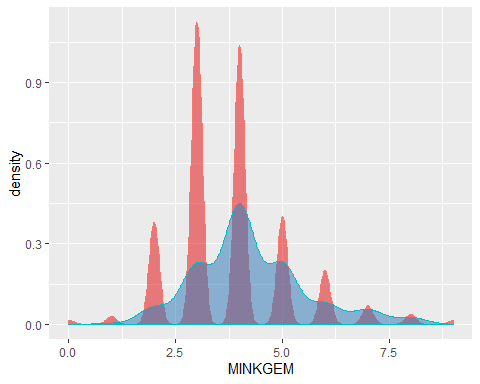
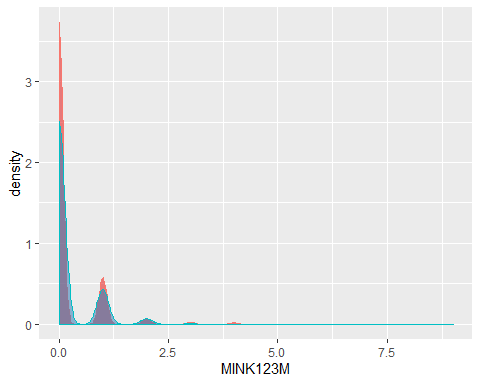
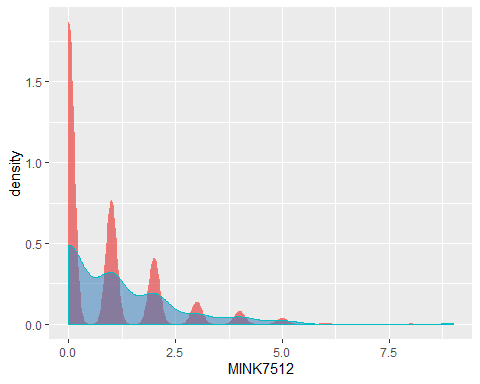
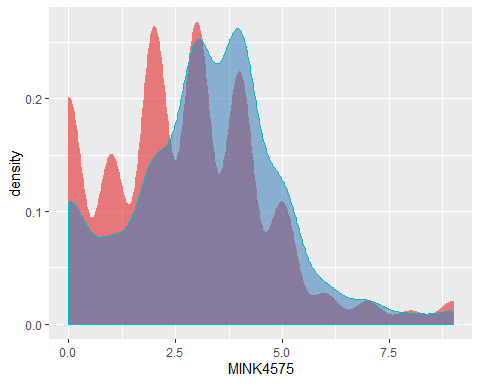
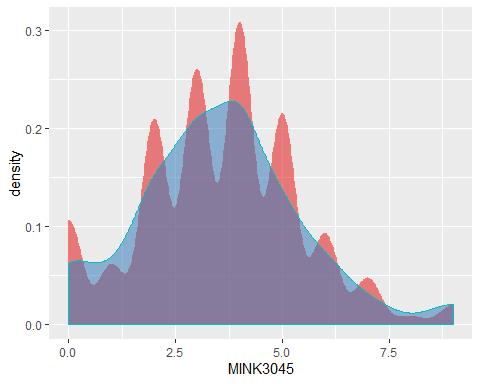
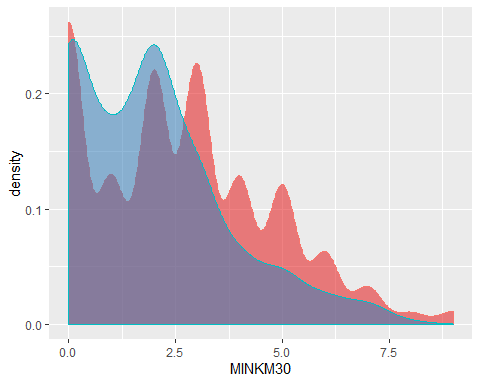
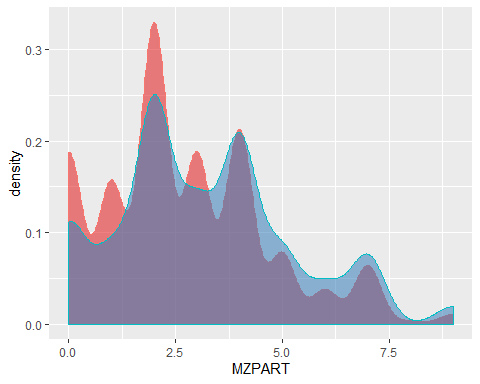
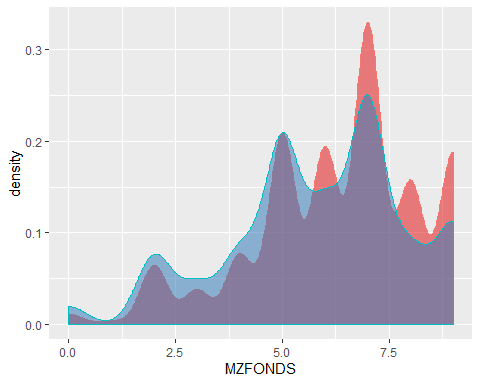
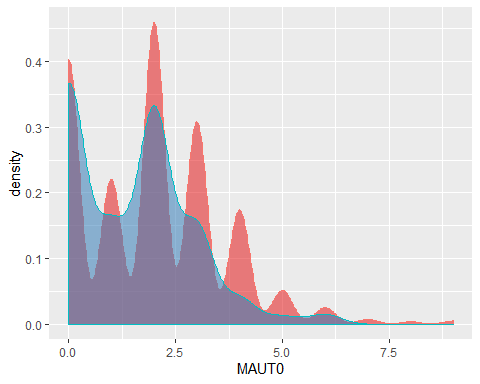
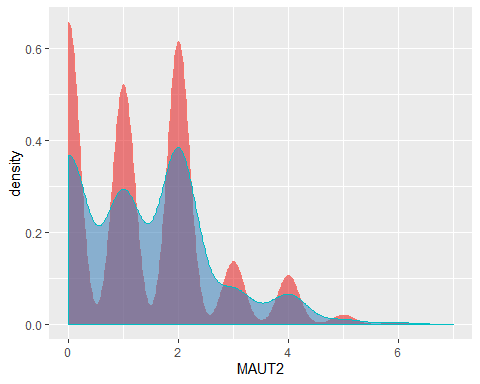
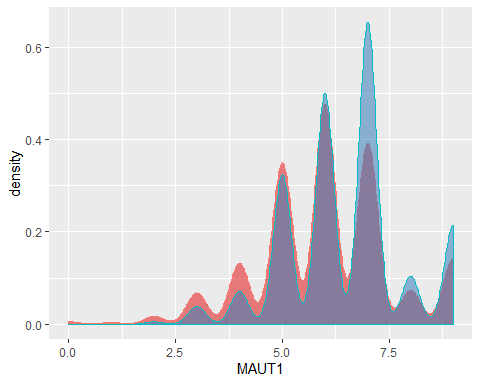
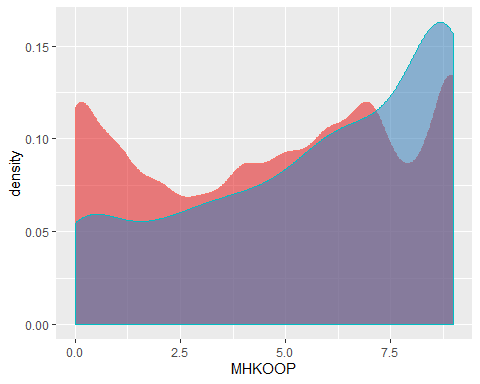
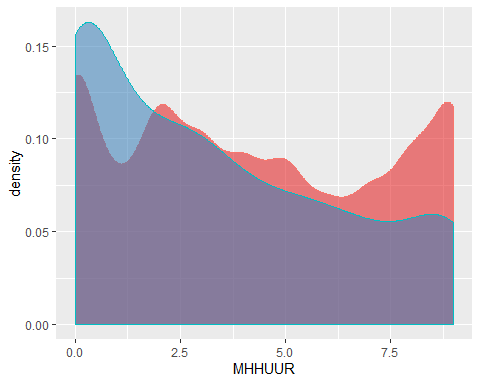
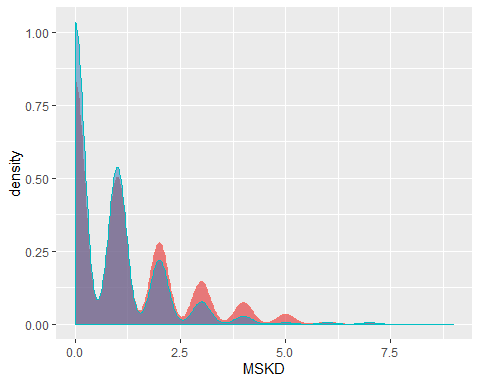
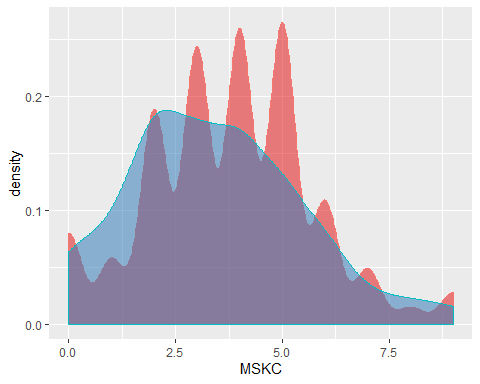
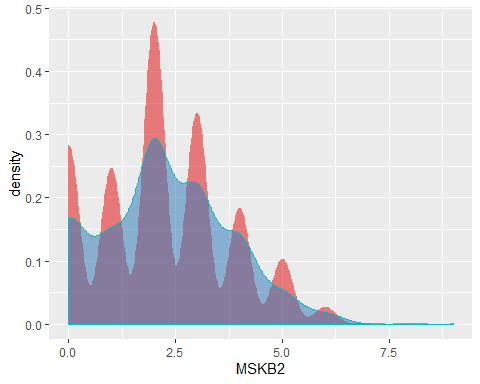
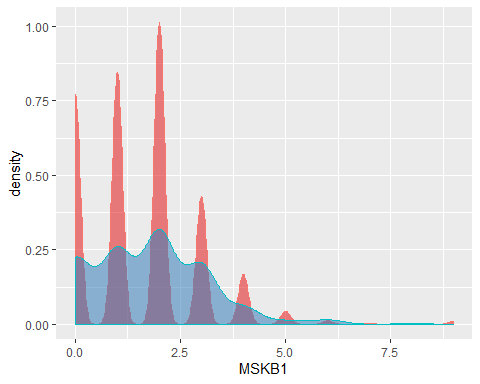
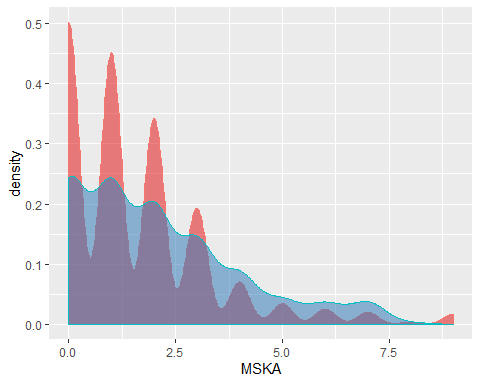
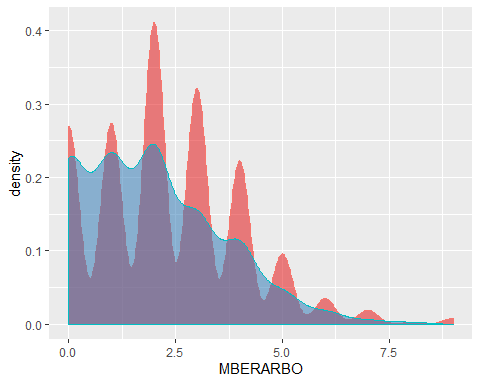
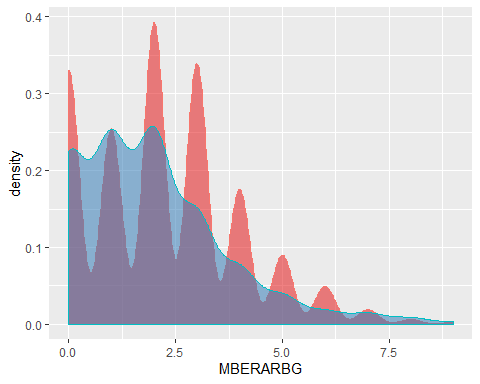
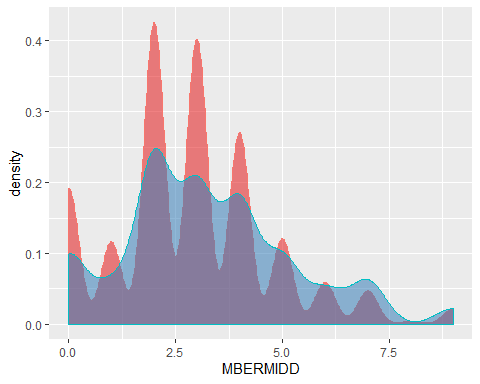
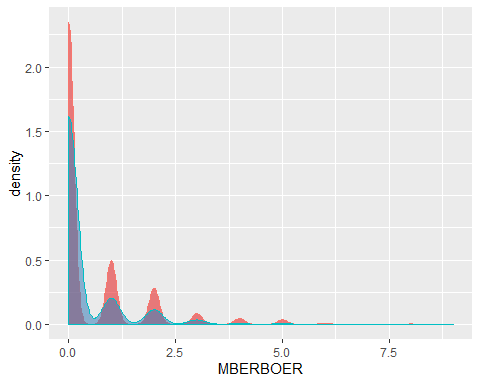
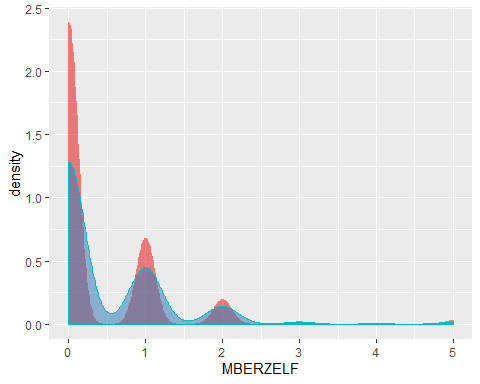
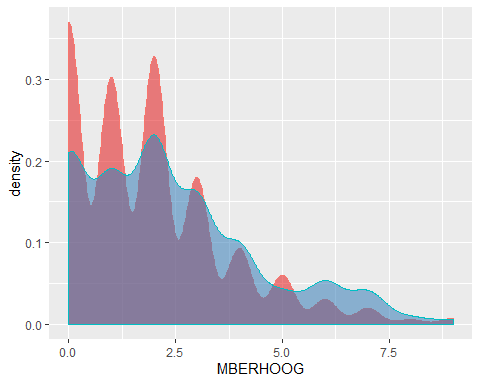
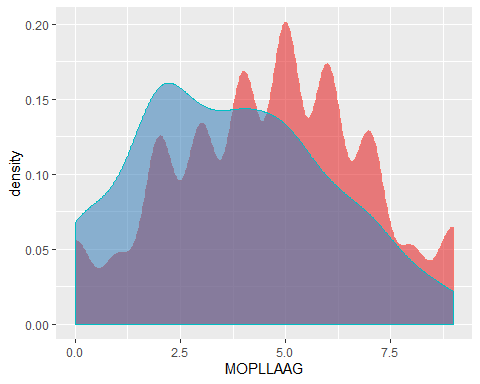
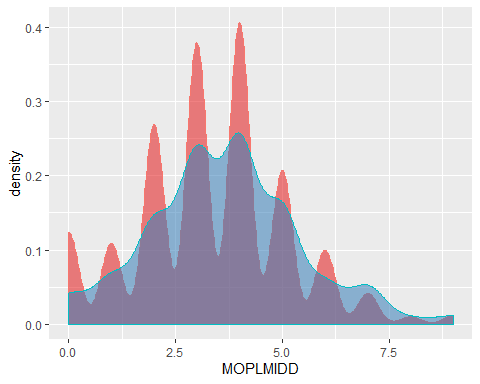
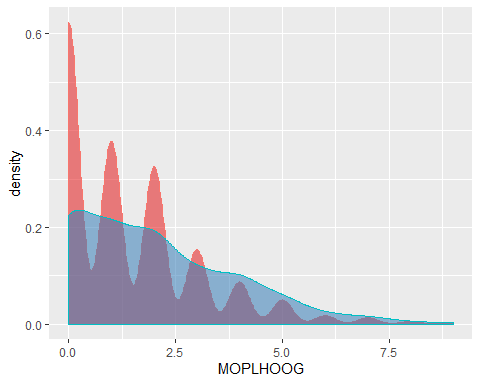
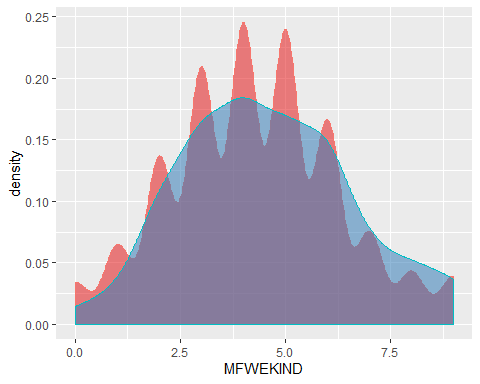
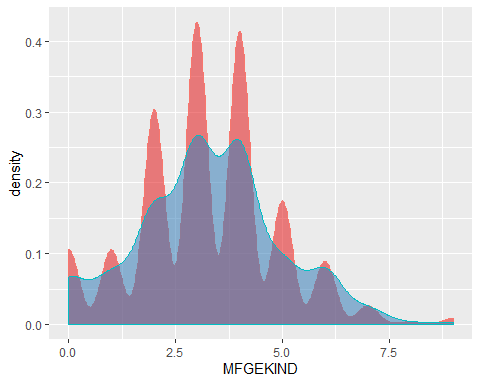
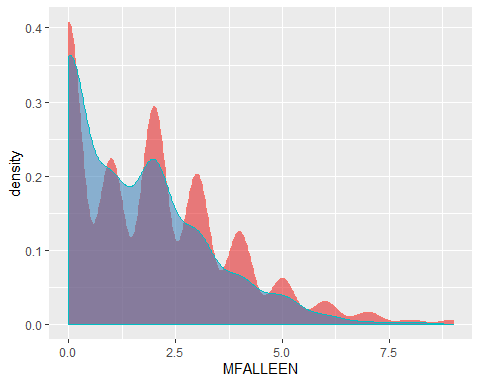
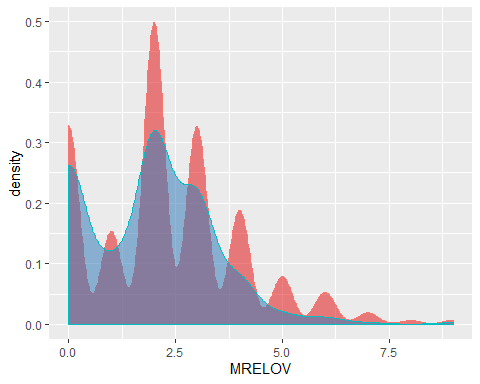
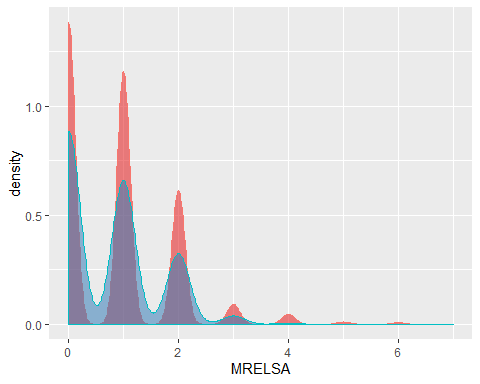
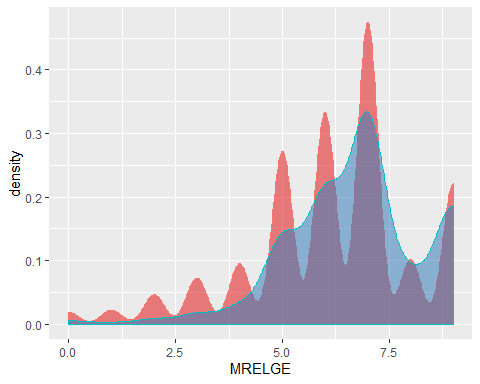
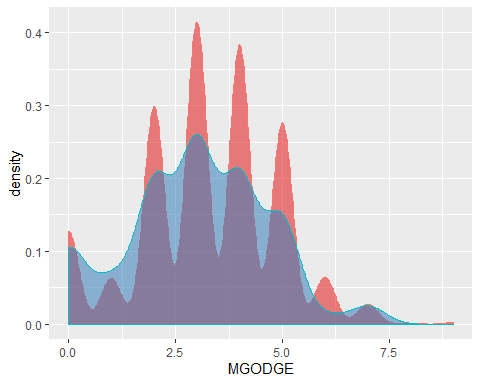
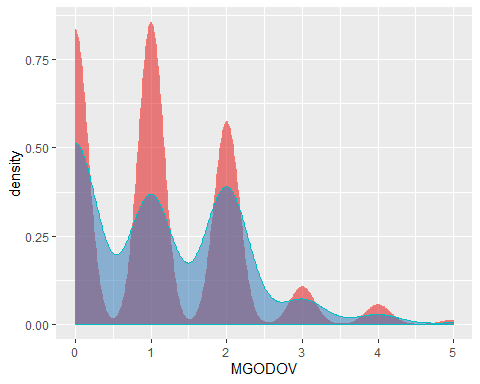
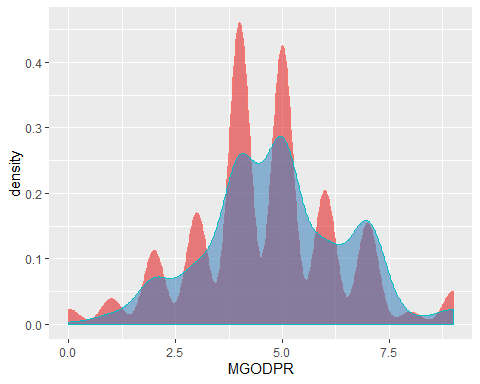
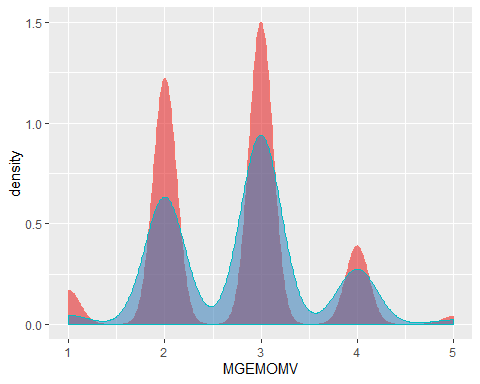
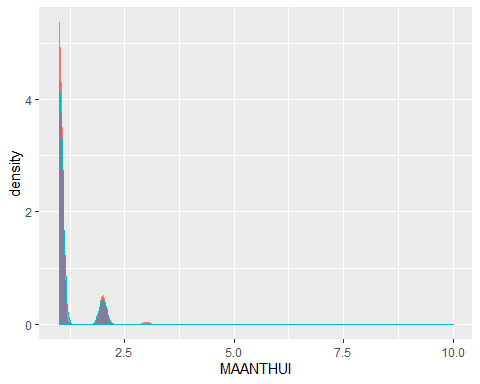
The dataset is unbalanced. There are only 348 cases that have purchased the mobile home insurance policy and 5474 that have not.

## Explanatory Variables

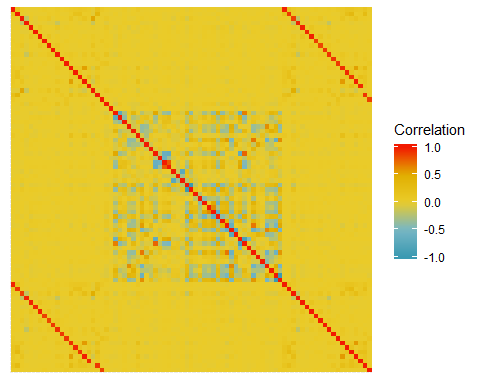
### Categorical



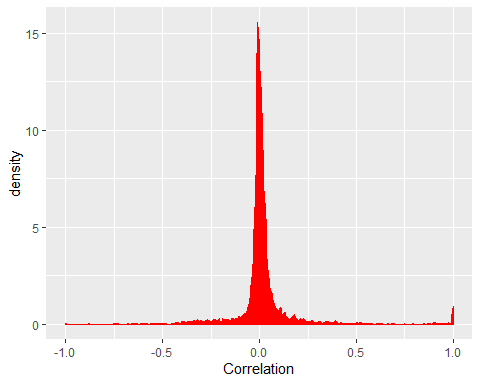
### Numeric



Here is a correlation heatmap. Although you can’t see the variable names, you get a general sense of how they are correlated:



Most of the correlation heatmap is yellow which is has a correlation coeffient of zero. Here’s a density plot of the correlation coefficients:



# Appendices

## Data Dictionary

|  |  |
| --- | --- |
| Name | Description |
| AAANHANG | Number of trailer policies |
| ABESAUT | Number of delivery van policies |
| ABRAND | Number of fire policies |
| ABROM | Number of moped policies |
| ABYSTAND | Number of social security insurance policies |
| AFIETS | Number of bicycle policies |
| AGEZONG | Number of family accidents insurance policies |
| AINBOED | Number of property insurance policies |
| ALEVEN | Number of life insurances |
| AMOTSCO | Number of motorcycle/scooter policies |
| APERSAUT | Number of car policies |
| APERSONG | Number of private accident insurance policies |
| APLEZIER | Number of boat policies |
| ATRACTOR | Number of tractor policies |
| AVRAAUT | Number of lorry policies |
| AWABEDR | Number of third party insurance (firms) |
| AWALAND | Number of third party insurane (agriculture) |
| AWAOREG | Number of disability insurance policies |
| AWAPART | Number of private third party insurance |
| AWERKT | Number of agricultural machines policies |
| AZEILPL | Number of surfboard policies |
| CARAVAN | Number of mobile home policy |
| MAANTHUI | Number of houses |
| MAUT0 | No car |
| MAUT1 | 1 car |
| MAUT2 | 2 cars |
| MBERARBG | Skilled labourers |
| MBERARBO | Unskilled labourers |
| MBERBOER | Farmer |
| MBERHOOG | High status |
| MBERMIDD | Middle management |
| MBERZELF | Entrepreneur |
| MFALLEEN | Singles |
| MFGEKIND | Household without children |
| MFWEKIND | Household with children |
| MGEMLEEF | Avg age |
| MGEMOMV | Avg size household |
| MGODGE | No religion |
| MGODOV | Other religion |
| MGODPR | Protestant |
| MGODRK | Roman catholic |
| MHHUUR | Rented house |
| MHKOOP | Home owners |
| MINK123M | Income >123.000 |
| MINK3045 | Income 30-45.000 |
| MINK4575 | Income 45-75.000 |
| MINK7512 | Income 75-122.000 |
| MINKGEM | Average income |
| MINKM30 | Income < 30.000 |
| MKOOPKLA | Purchasing power class |
| MOPLHOOG | High level education |
| MOPLLAAG | Lower level education |
| MOPLMIDD | Medium level education |
| MOSHOOFD | Customer main type |
| MOSTYPE | Customer Subtype |
| MRELGE | Married |
| MRELOV | Other relation |
| MRELSA | Living together |
| MSKA | Social class A |
| MSKB1 | Social class B1 |
| MSKB2 | Social class B2 |
| MSKC | Social class C |
| MSKD | Social class D |
| MZFONDS | National Health Service |
| MZPART | Private health insurance |
| PAANHANG | Contribution trailer policies |
| PBESAUT | Contribution delivery van policies |
| PBRAND | Contribution fire policies |
| PBROM | Contribution moped policies |
| PBYSTAND | Contribution social security insurance policies |
| PFIETS | Contribution bicycle policies |
| PGEZONG | Contribution family accidents insurance policies |
| PINBOED | Contribution property insurance policies |
| PLEVEN | Contribution life insurances |
| PMOTSCO | Contribution motorcycle/scooter policies |
| PPERSAUT | Contribution car policies |
| PPERSONG | Contribution private accident insurance policies |
| PPLEZIER | Contribution boat policies |
| PTRACTOR | Contribution tractor policies |
| PVRAAUT | Contribution lorry policies |
| PWAAND | Contribution third party insurane (agriculture) |
| PWABEDR | Contribution third party insurance (firms) |
| PWAOREG | Contribution disability insurance policies |
| PWAPART | Contribution private third party insurance |
| PWERKT | Contribution agricultural machines policies |
| PZEILPL | Contribution surfboard policies |

## References

Elkan, Charles. 2000. “CoIL Challenge 2000 Entry.” <http://liacs.leidenuniv.nl/~puttenpwhvander/library/cc2000/ELKANP~1.pdf>.

Putten, Peter, Michel Ruiter, and Maarten Someren. 2000. “CoIL Challenge 2000 Tasks and Results: Predicting and Explaining Caravan Policy Ownership.” <http://liacs.leidenuniv.nl/~puttenpwhvander/library/cc2000/PUTTEN~1.pdf>.