Analysis of Different Predictive Models for Prediction of Aromatic Rings in SRC Kinase Inhibitors

Keerthi Krishnan

Introduction and Project Description

Protein kinases are key enzymes in many signal transduction pathways, and play a crucial role in cellular proliferation, differentiation, and various cell regulatory processes and the human genome encodes 538 kinase genes, which make up nearly 2% of all genes[1,2,3]. Due to their pivotal role in signal transduction and cell cycle pathways, they are known to be the cause of many common diseases such as cancer[4]. Around 400 diseases are associated with protein kinases directly or indirectly. protein kinase inhibitors (PKIs) have emerged as a subject of great theoretical importance and therapeutic value [5]. Based on their binding modes with targeted protein kinases, small molecule PKIs can be classified into Type 1, 2, and 3 inhibitors[6]. A Type 1 inhibitor is defined as a small molecule that binds to the active conformation of a kinase in the ATP pocket; the Type 2 inhibitor binds to an inactive (usually DFG-OUT) conformation of a kinase; and the Type 3 inhibitor binds next to the ATP-binding pocket allosterically and is a non-ATP competitive inhibitor.[7]

Several molecular descriptors, including molecular weight, number of donor atoms for hydrogen bonds, number of acceptor atoms for hydrogen bonds, cLogP, number of rotatable bonds and topological polar surface area are important descriptors of protein kinases because of their association with established roles for "drug likeness" from the perspective of bioavailability[8]. Other molecular descriptors such as the number of aromatic rings, aromatic ratio, and fraction of sp3 carbon atoms, along with hydrogen bond counts, are considered to be important from the perspective of drug–protein binding affinity.As in all ligand–protein complexes, the molecular recognition between PKIs and their target protein kinases are achieved by non-bonded interactions[9].

SRC Kinase Inhibitors are a type of protein kinase inhibitors and act as potential drug candidates for cancer proteins. They play an vital role in suppressing tumor growth and oncogenesis in tumor proteins. A distinction factor of SRC Kinase Inhibitors and most other inhibitor molecules is their existence is their ring count, acting as a key feature or characteristic for these molecules. Potentially, the ring count can affect binding modes of these inhibitors to different cancer proteins and in turn, affect the potential inhibition of cancer proteins. In my research lab, we work in generative modeling and design of SRC Kinase Inhibitors where one of the key evaluating measures of generated molecules is the ring count. Therefore, I plan on predicting aromatic ring count using purely chemical feature based Machine Learning models. There were 2 objectives I wanted to achieve in this project:

1. Utilizing Machine Learning to build a model that can predict the number of rings of potential SRC Kinase Inhibitor molecules based on chemical properties and salient features.
2. Comparative Analysis of Different Machine Learning Models to see which model predicts accurately and provides the best fit for the data.

For this project I implemented 4 different predictive models: Linear Regression Model, Log-Linear Regression Model, Random Forest Model, and a Neural Network Model. While building each model, I will explain the use of implementation and the analysis.

The 3 phases of this project will include pre-processing, model building, and model performance analysis. I will first describe pre-processing mechanisms taken, then will go into model building and performance analysis for each model, and then will conclude with my findings and further remarks.

Data Description and Pre-Processing Mechanism

The data that I used consists of the smile strings of the molecules and ~20 chemical attributes for each molecule, which also include drug-likeliness properties. In total, there are ~3000 instances.

Smile strings act as a form to describe the molecular formula of your molecule. You can think of them as the makeup of your molecule and can give clues on how the molecule will be structurally and property-wise.

There are 20 aggregate chemical features, in which some features describe the basic characteristics of the molecule such as weight, number of rotatable bonds, hba values, hbd values, etc. However, there are also features that describe drug-likeliness characteristics of the molecule such as logP, SAS, and QED values. Having a mix of different chemical characteristics and features is always good for a model as it creates more variance in the dataset and can give different perspectives in understanding model fitting.

Our outcome or predictor variable will be rings. This variable represents the number of aromatic rings in a molecule.

Some of the pre-processing will include getting rid of unnecessary co-variates, getting rid of any missing values, and performing some initial visualization of the data to understand how each feature looks and its correlation to other features involved.

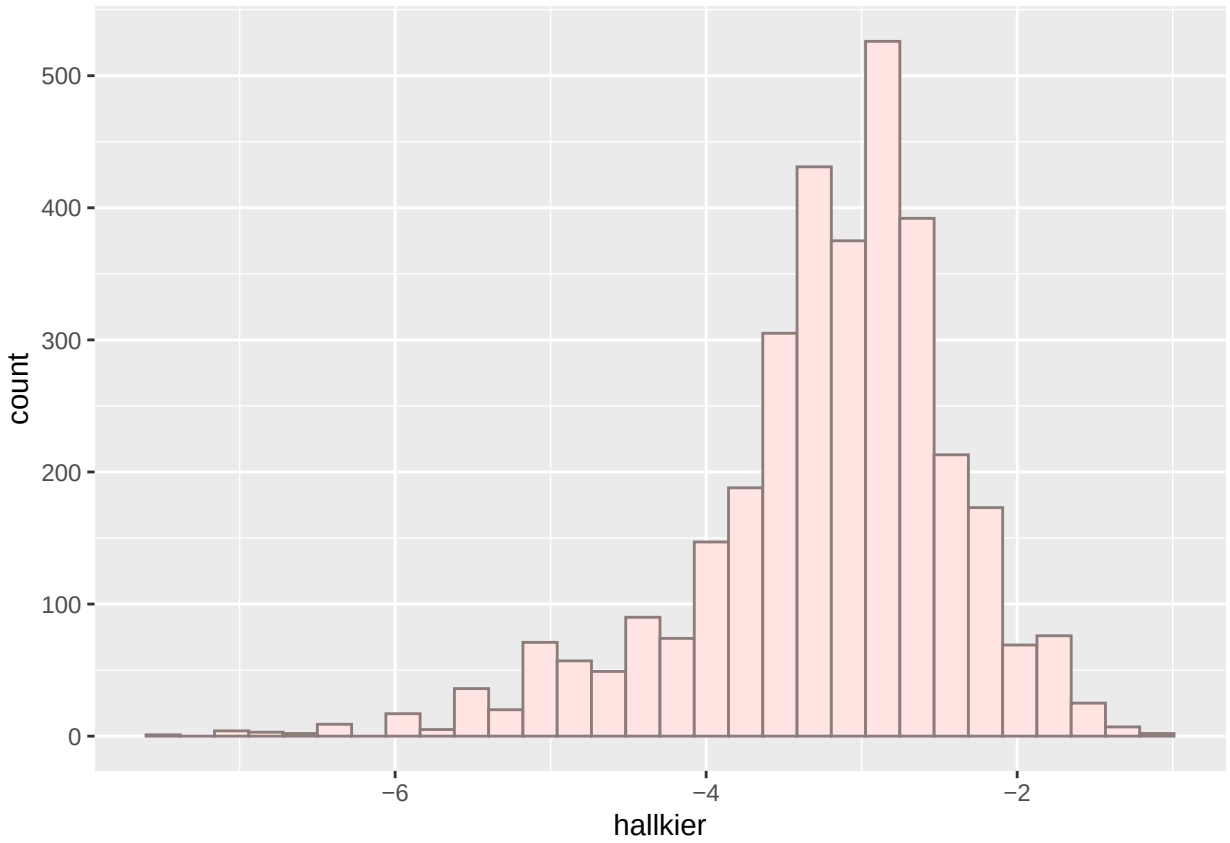
We will look at 3 data visualizations:

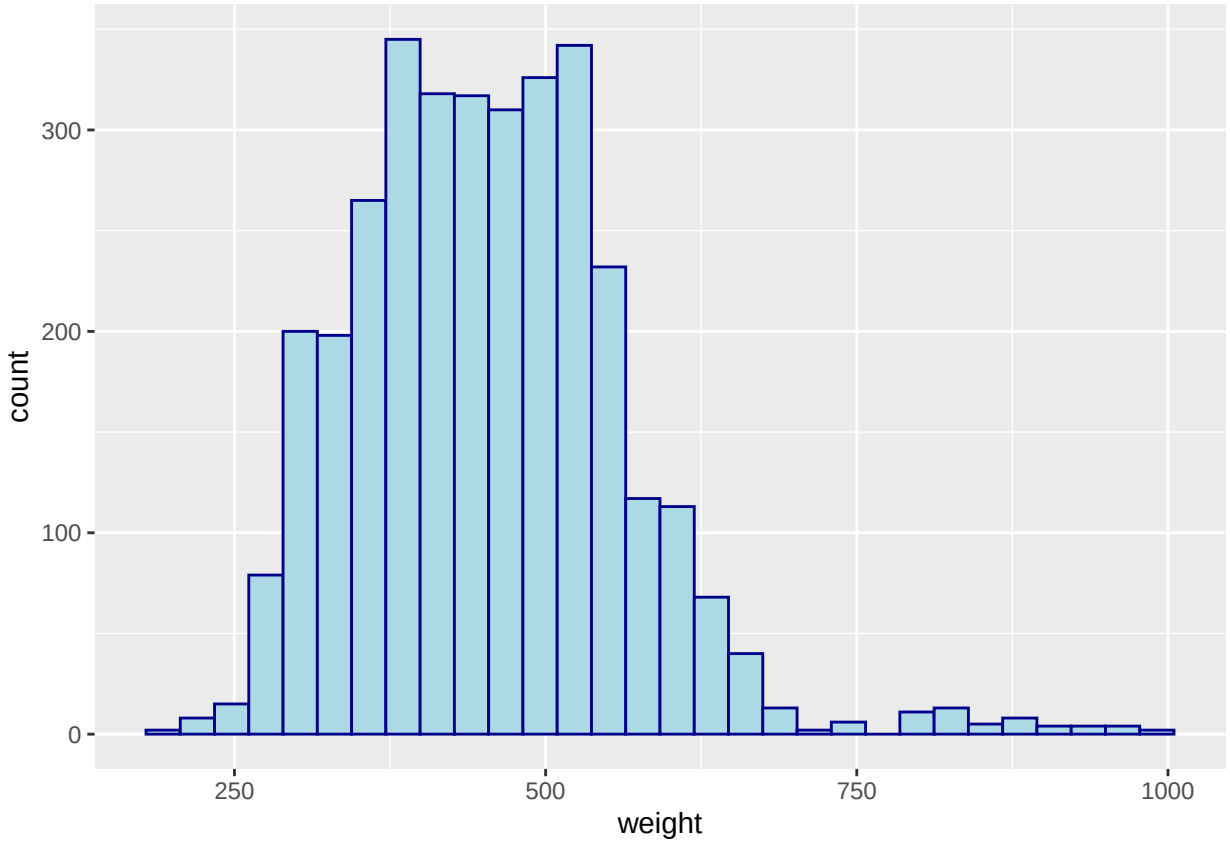
1. Histogram Distribution of Variables
2. Scatterplot Visualization of Co-Variates to Outcome Variable
3. Correlation Matrix Plot of Data Features

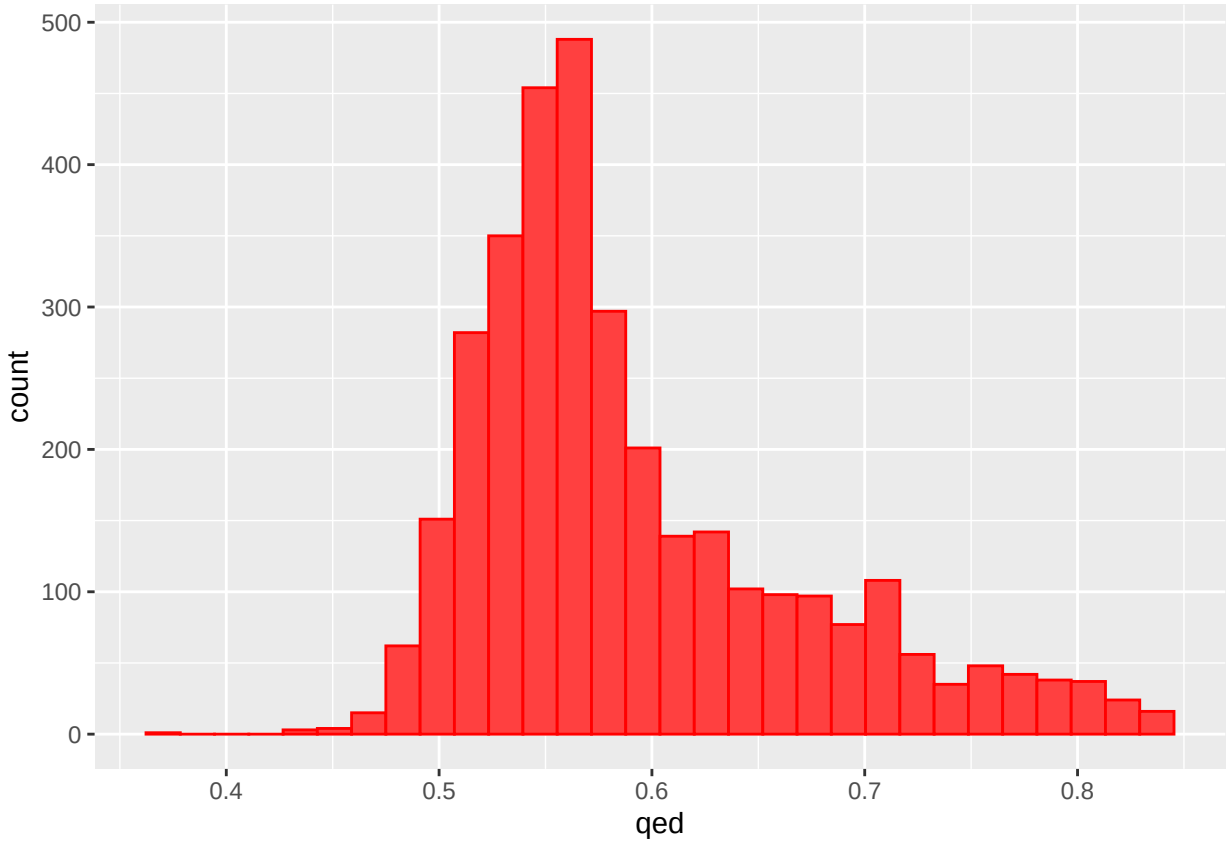
Analysis of Variables, Distributions, and Correlations- Visualization

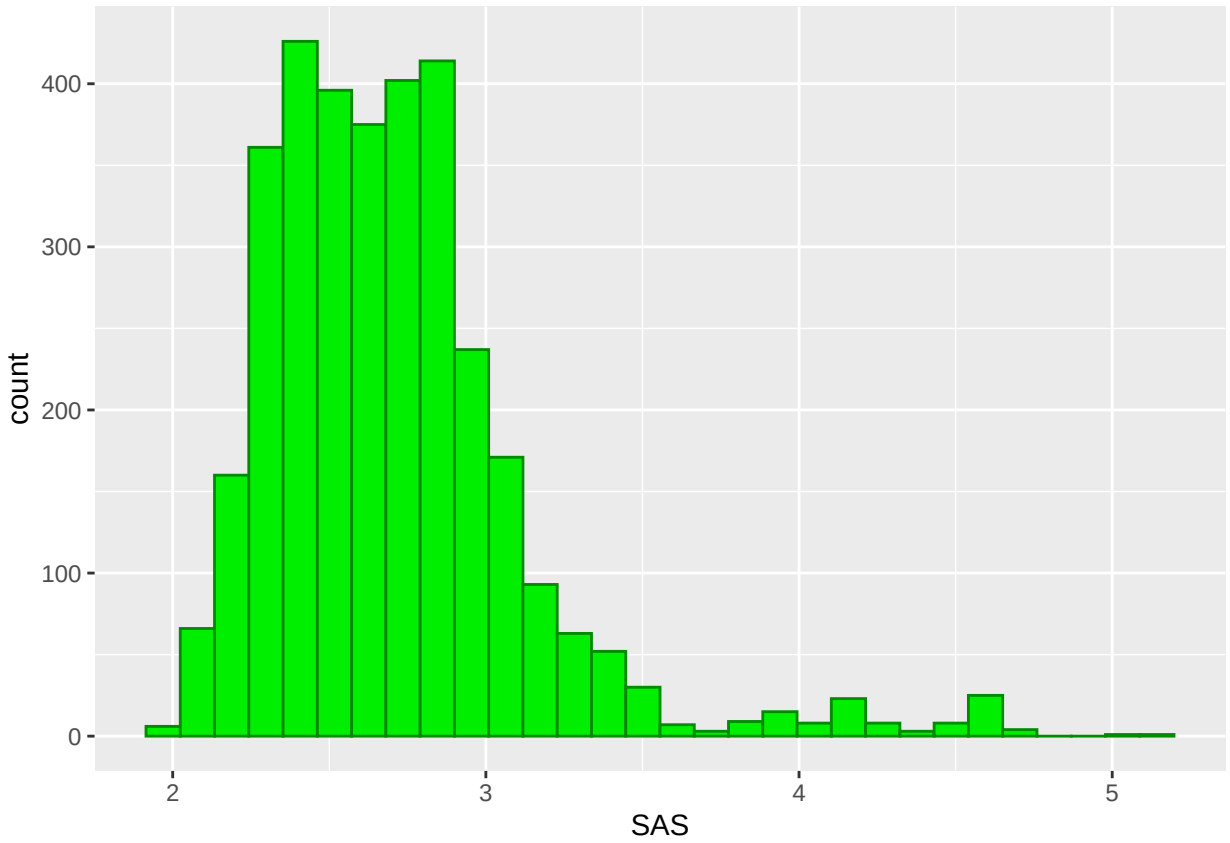
Histogram Visualization of Variables

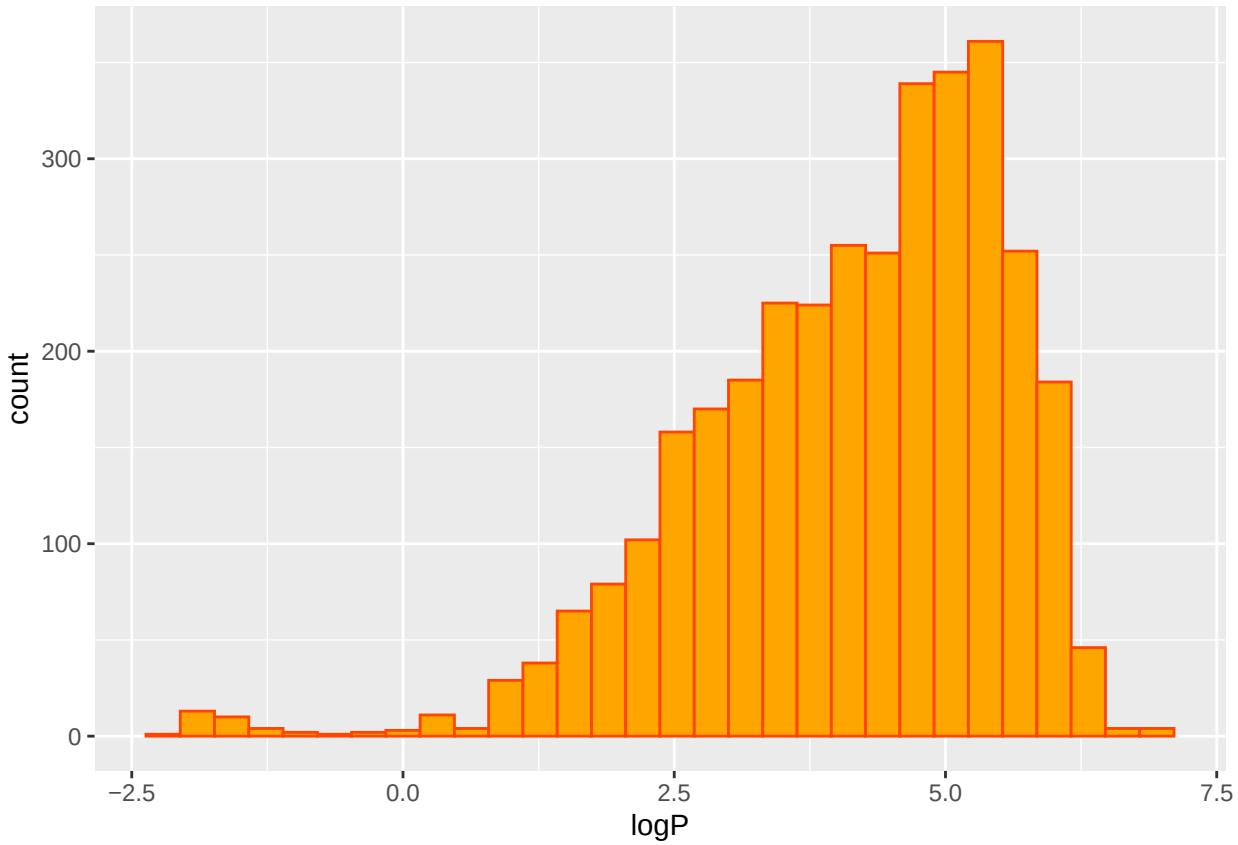
We use the histograms to understand how the data is distributed for each feature. This can potentially provide insight into the way we want to build the model and the importance of each feature.

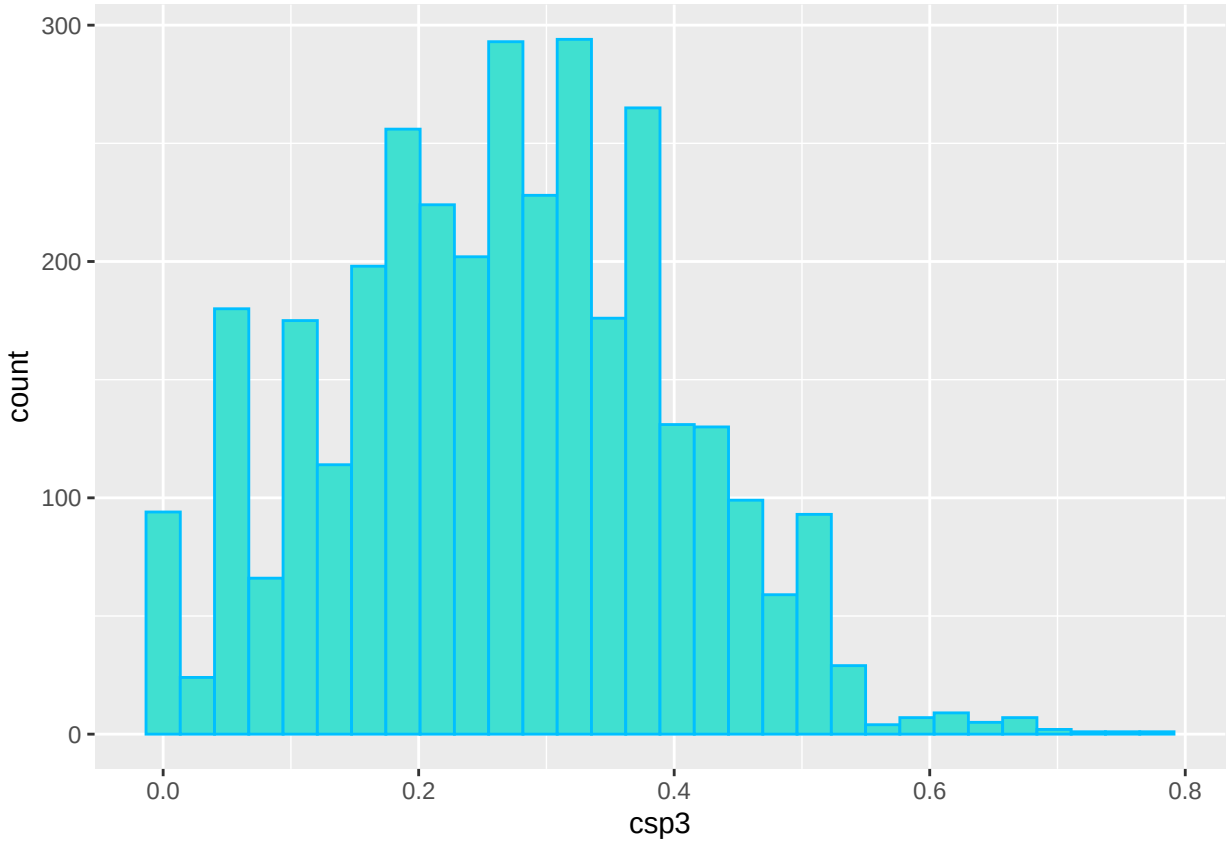

500
400
300
200
100
0
count
−6
−4
−2
hallkier

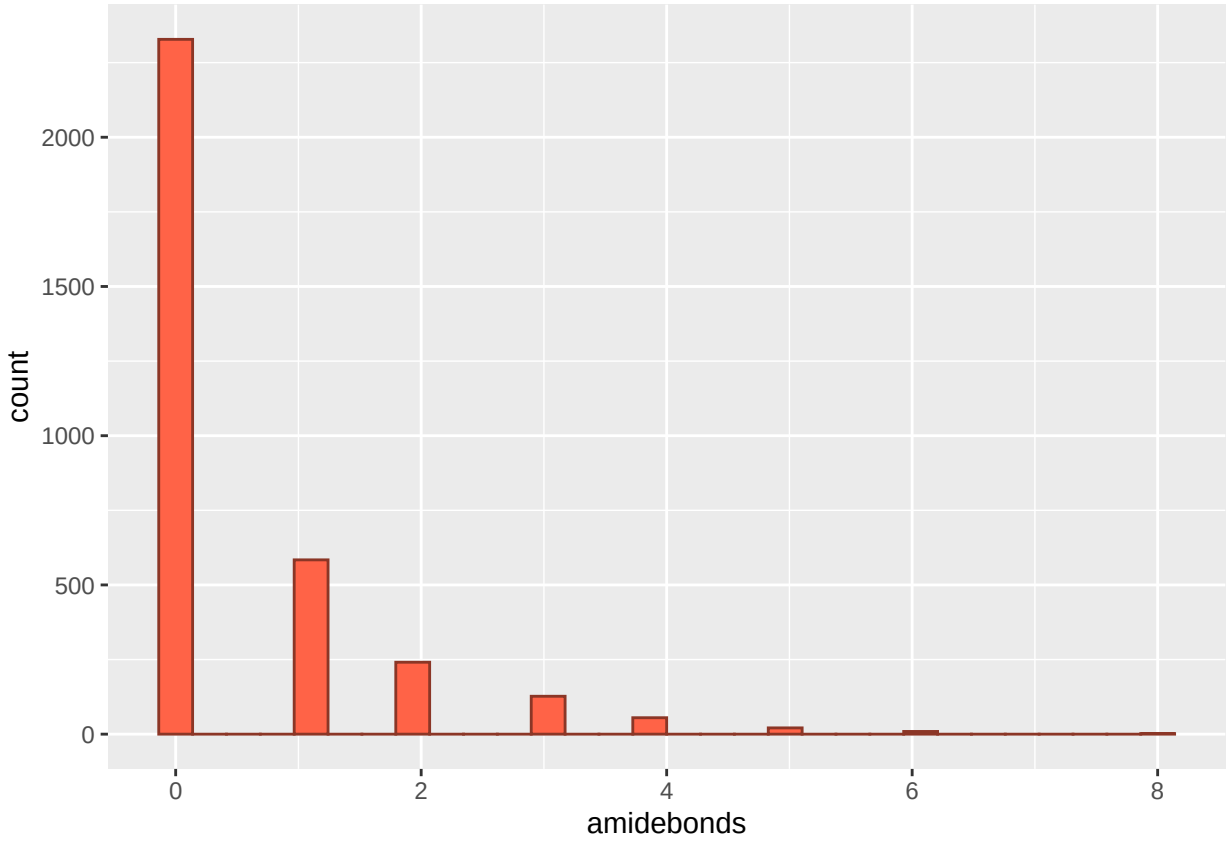

300
200
100
0
count
250
500
750
1000
weight

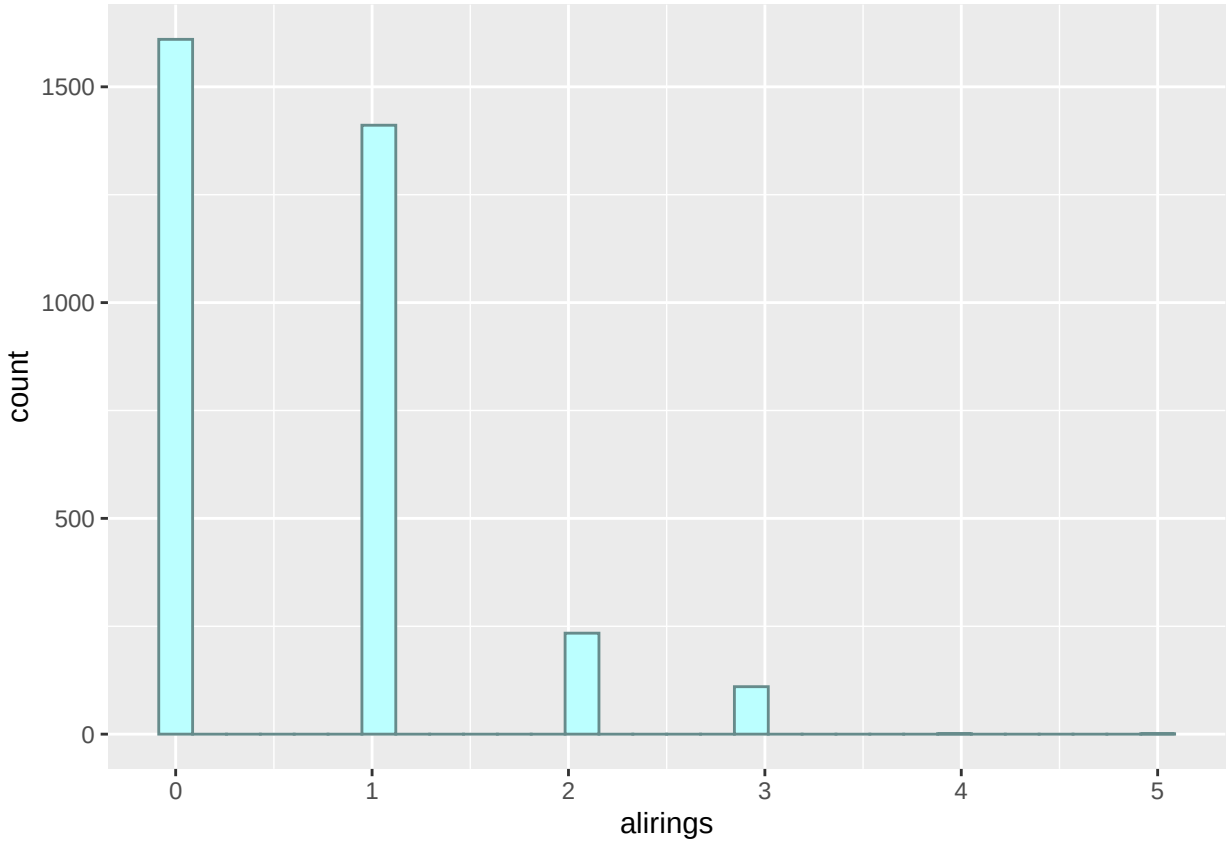

500
400
300
200
100
0
count
0.4
0.5
0.6
0.7
0.8
qed

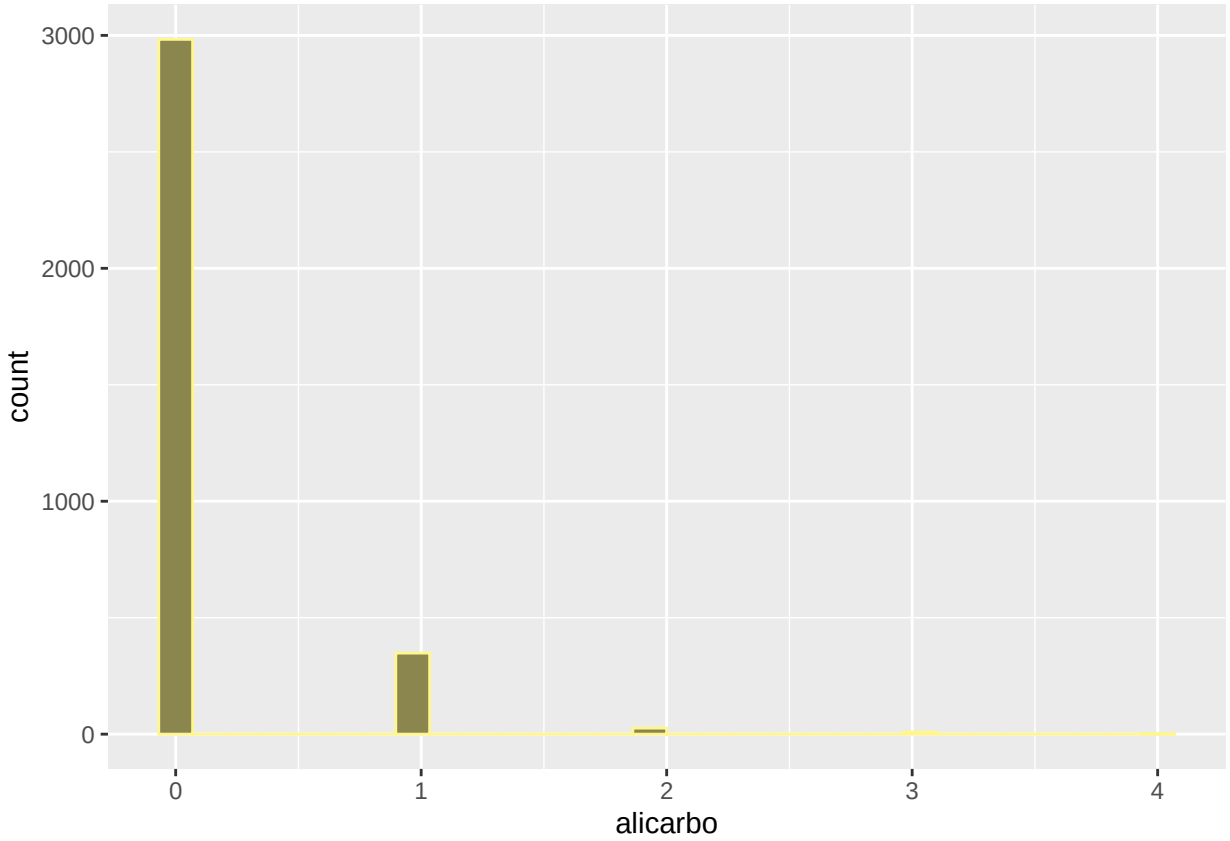

400
300
200
100
0
count
2
3
4
5
SAS

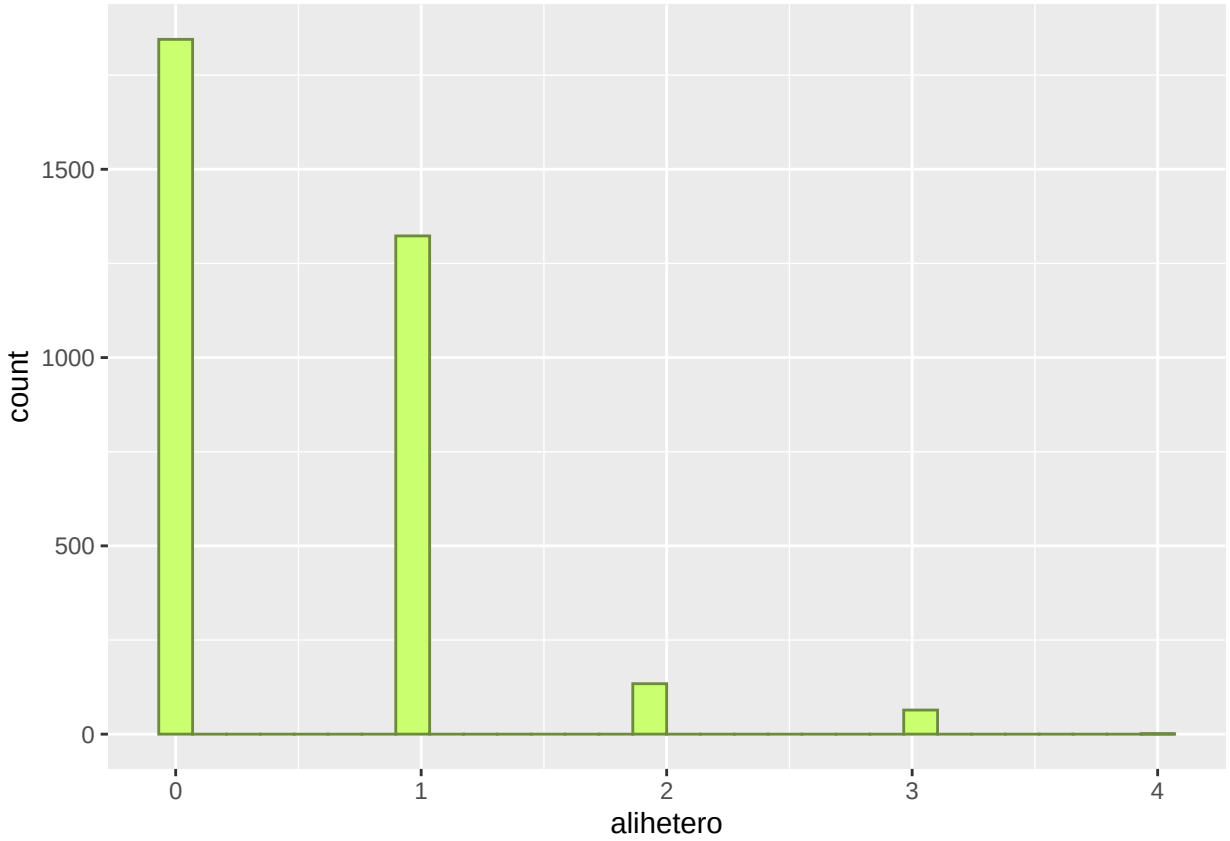

300
200
100
0
count
−2.5
0.0
2.5
5.0
7.5
logP

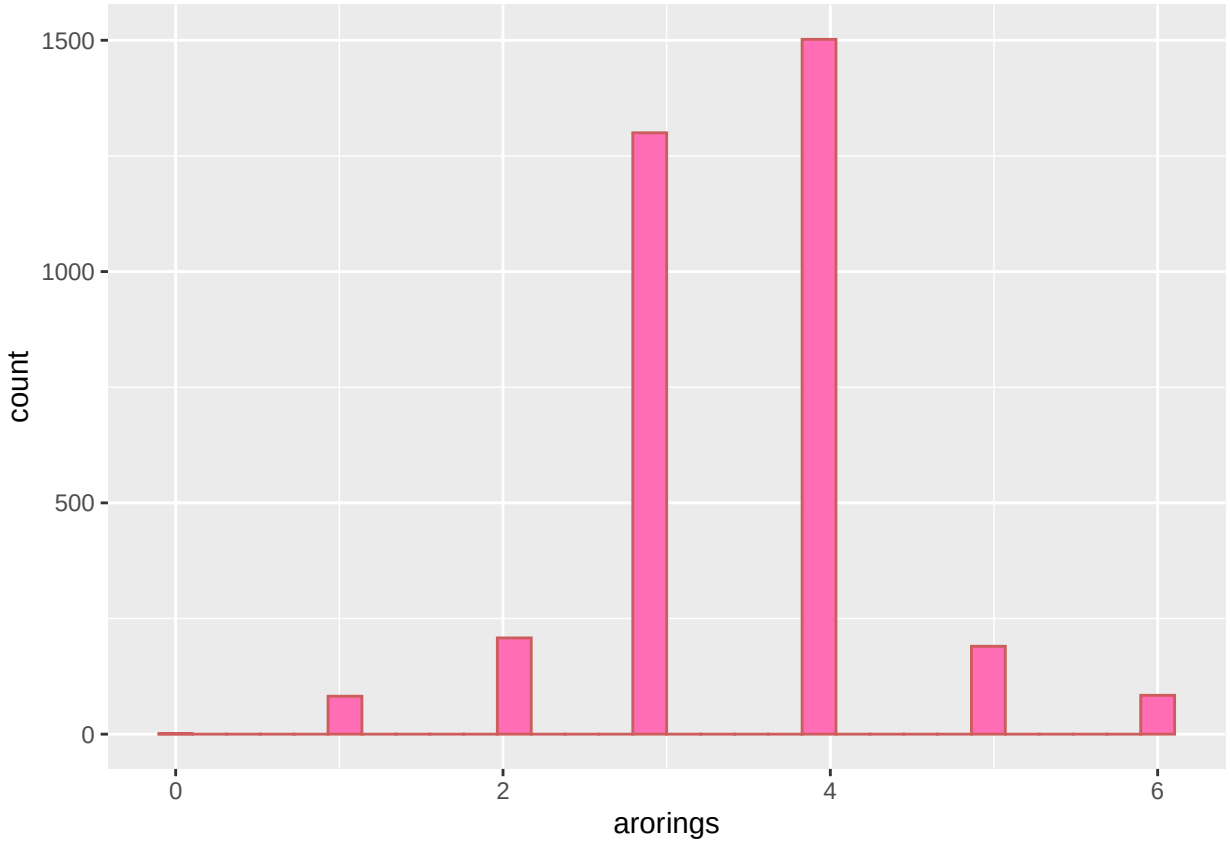

300
200
100
0
count
0.0
0.2
0.4
0.6
0.8
csp3

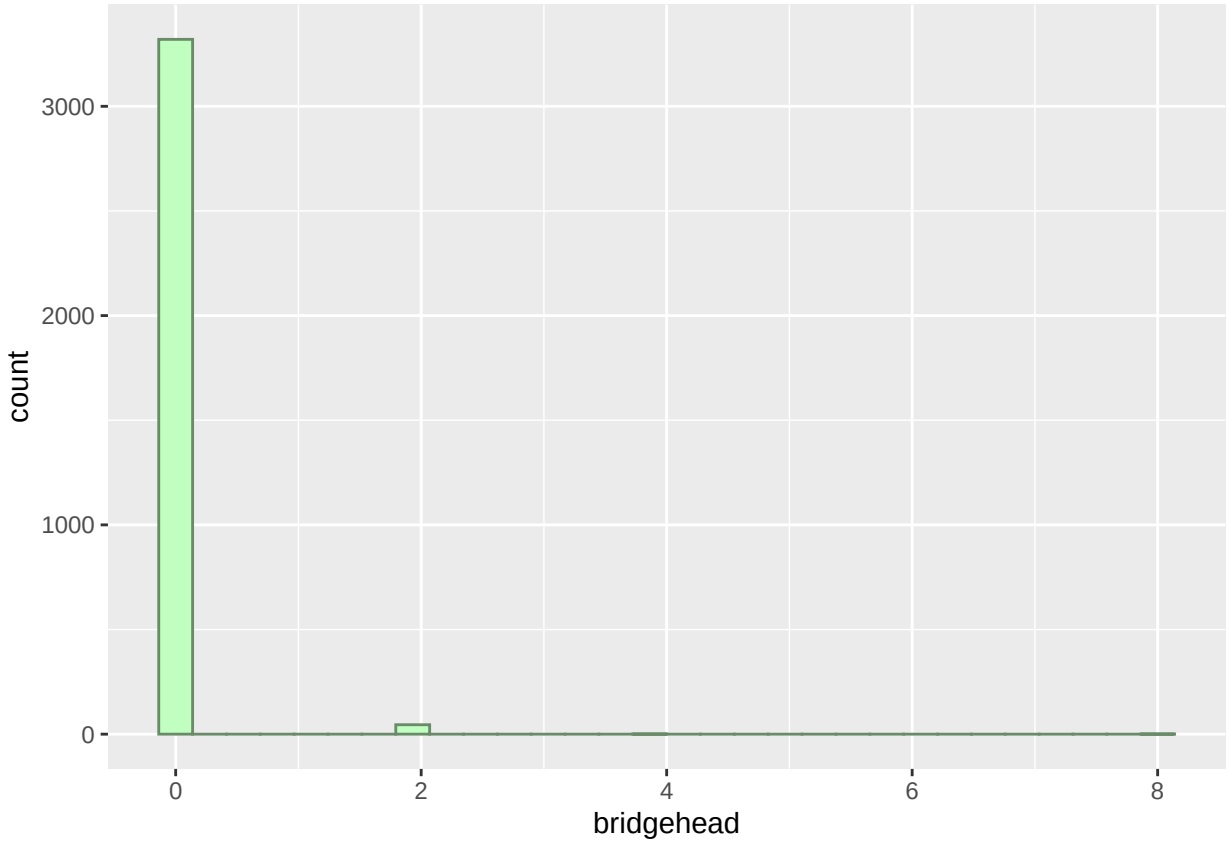

count
2000
1500
1000
500
0
0
2
4
6
8
amidebonds

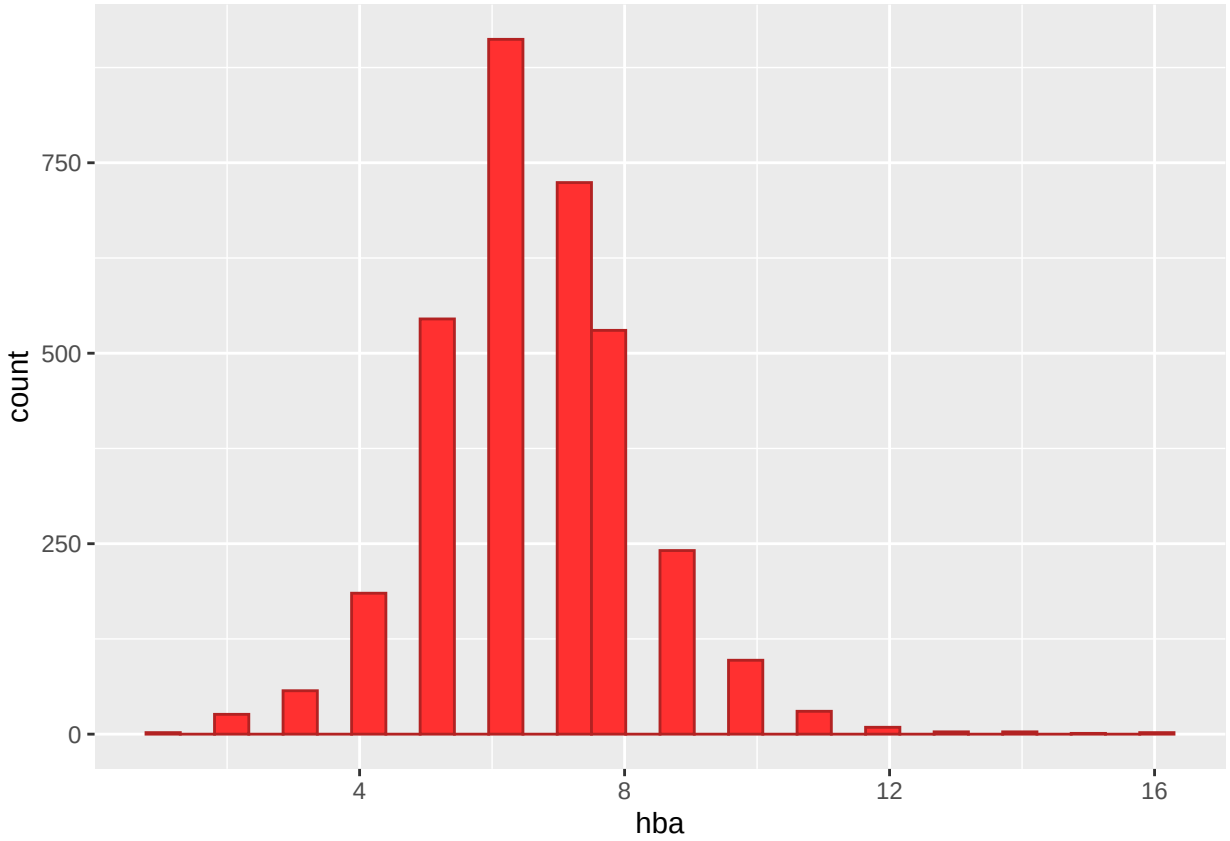

1500
1000
500
0
count
0
1
2
3
4
5
alirings

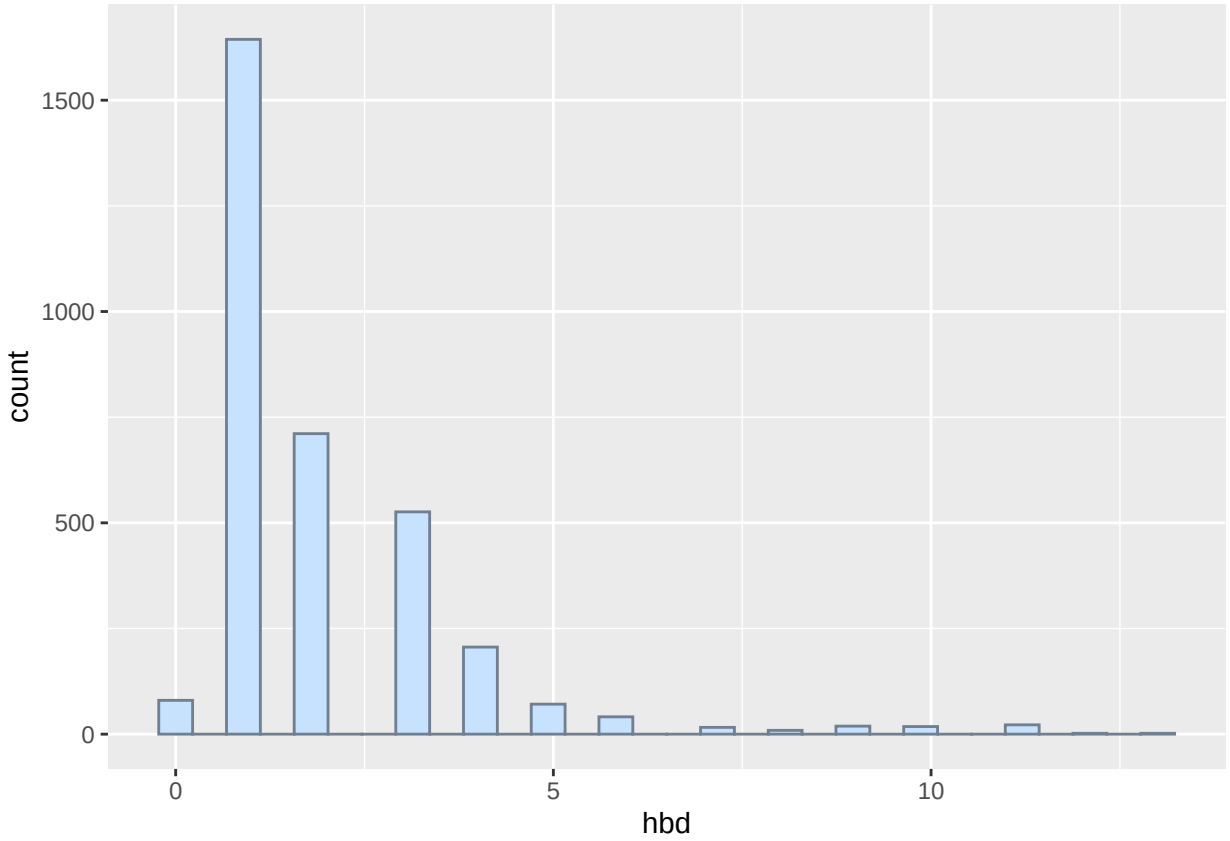

3000
2000
1000
0
count
0
1
2
3
4
alicarbo

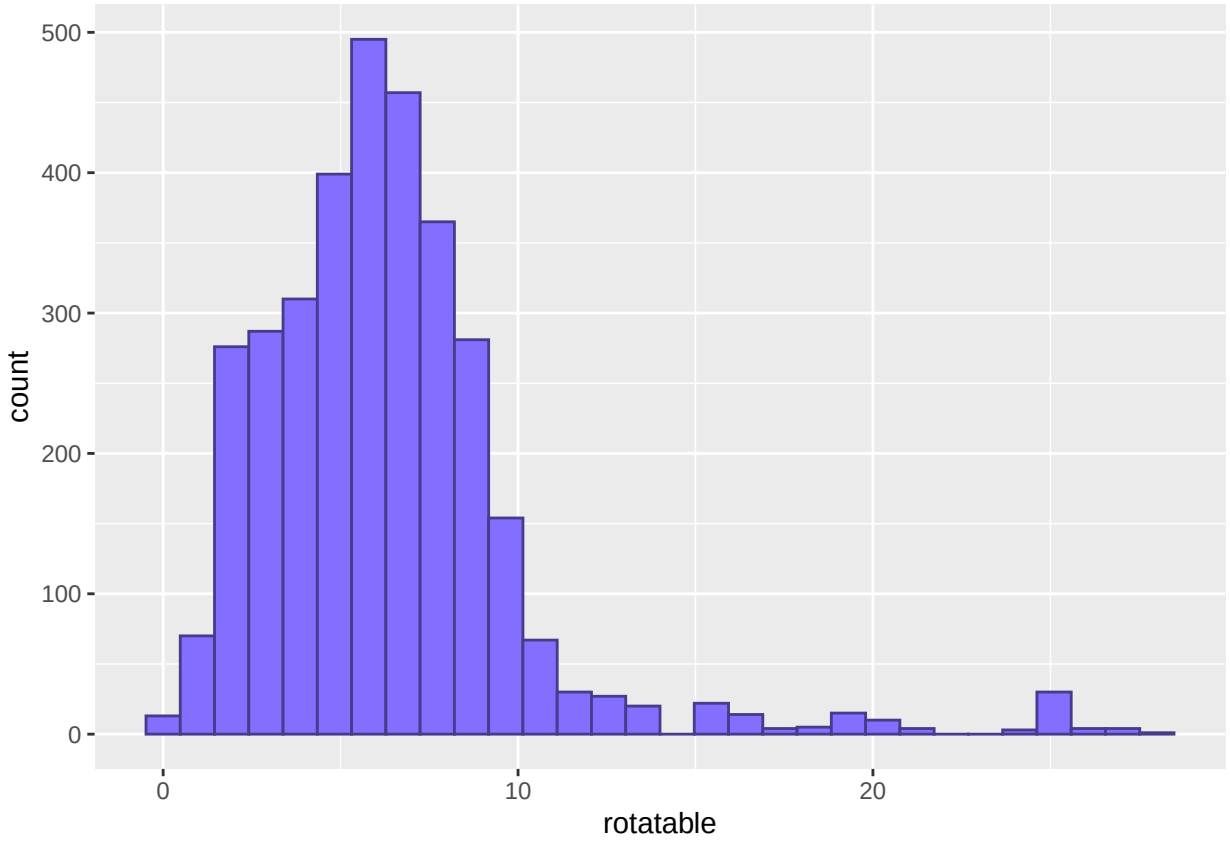

count
1500
1000
500
0
0
1
2
3
4
alihetero

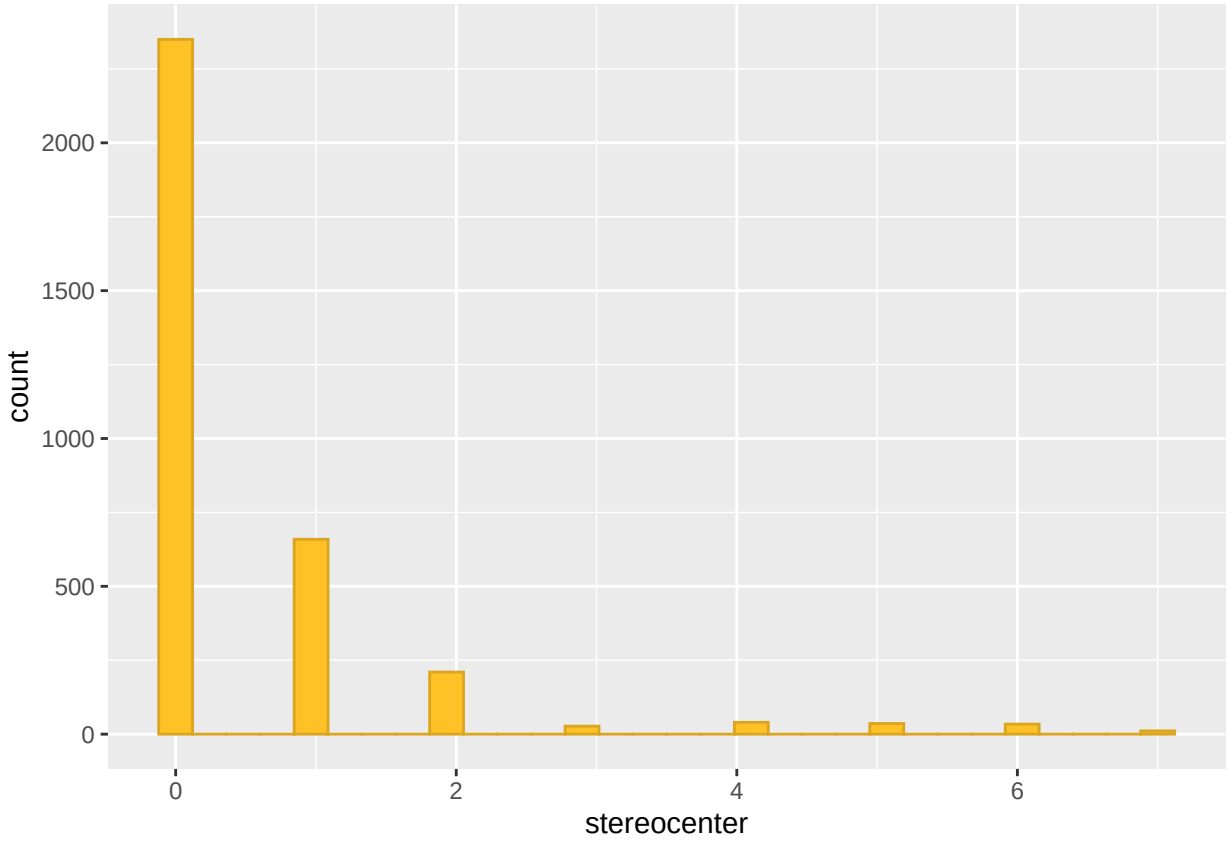

1500
1000
500
0
count
0
2
4
6
arorings

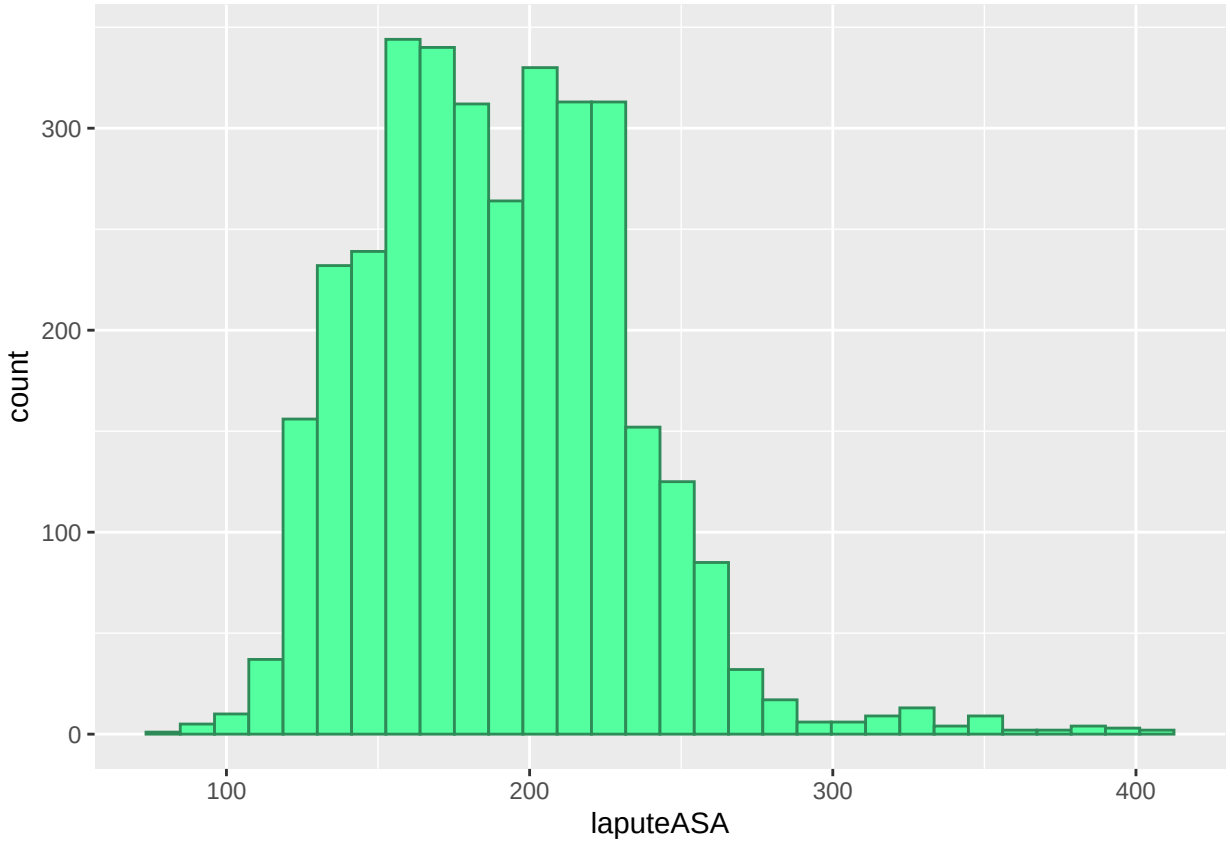

3000
2000
1000
0
count
0
2
4
6
8
bridgehead

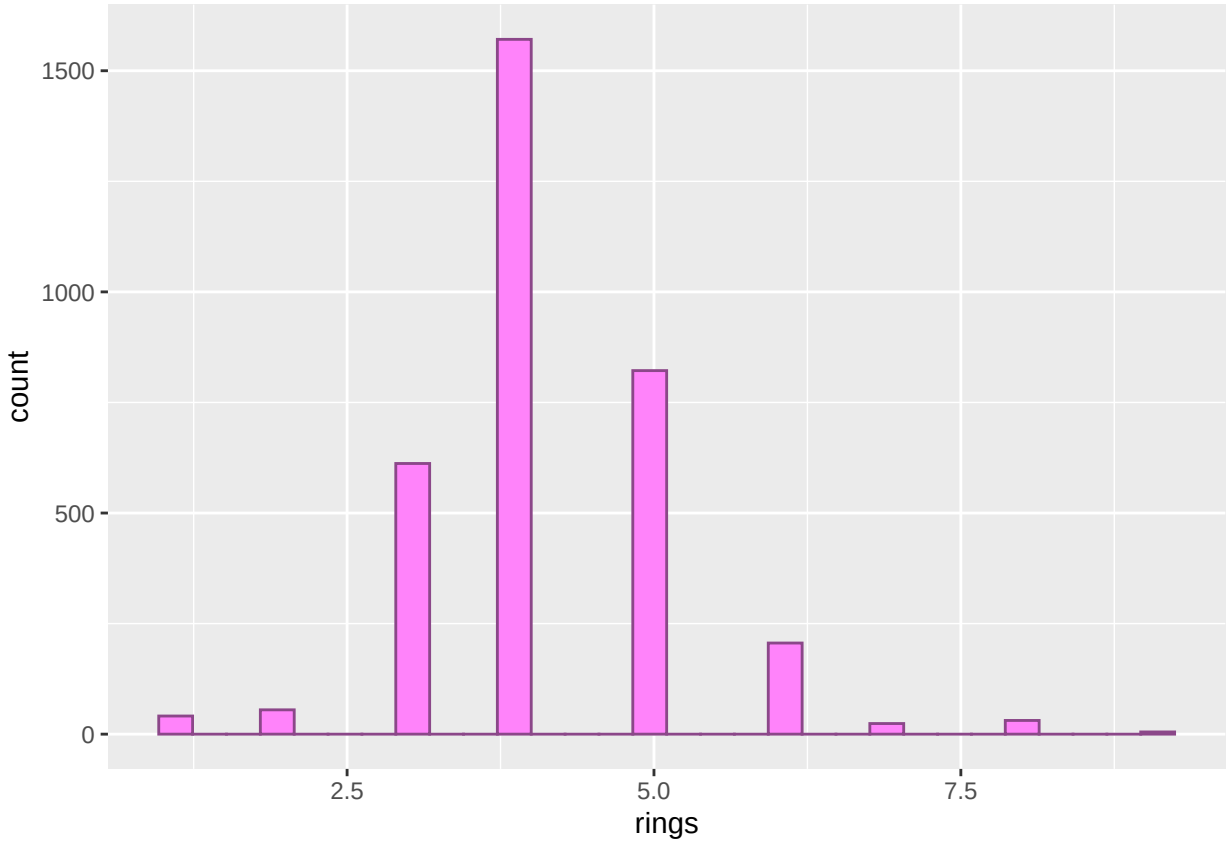

750
500
250
0
count
4
8
12
16
hba


1500
1000
500
0
count
0
5
10
hbd


500
400
300
200
100
0
count
0
10
20
rotatable

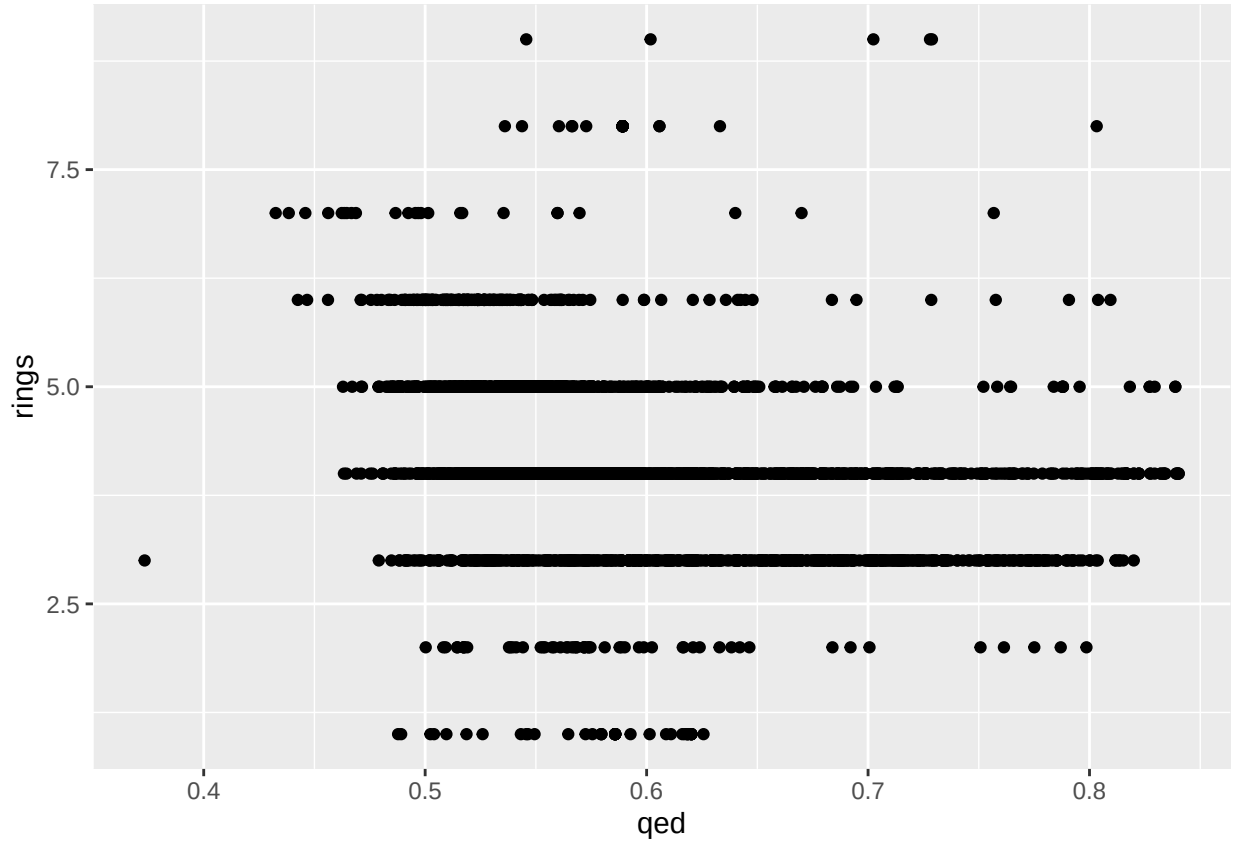

count
2000
1500
1000
500
0
0
2
4
6
stereocenter

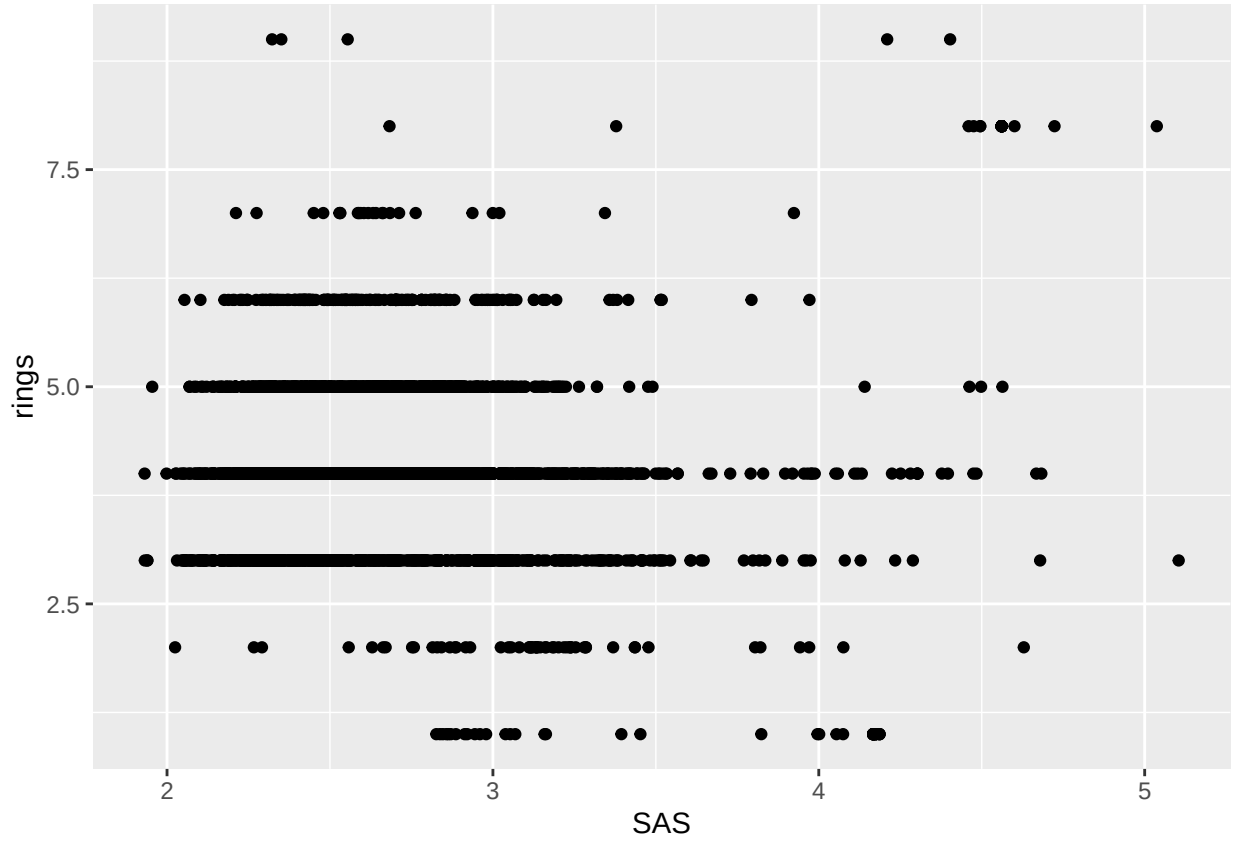

300
200
100
0
count
100
200
300
400
laputeASA

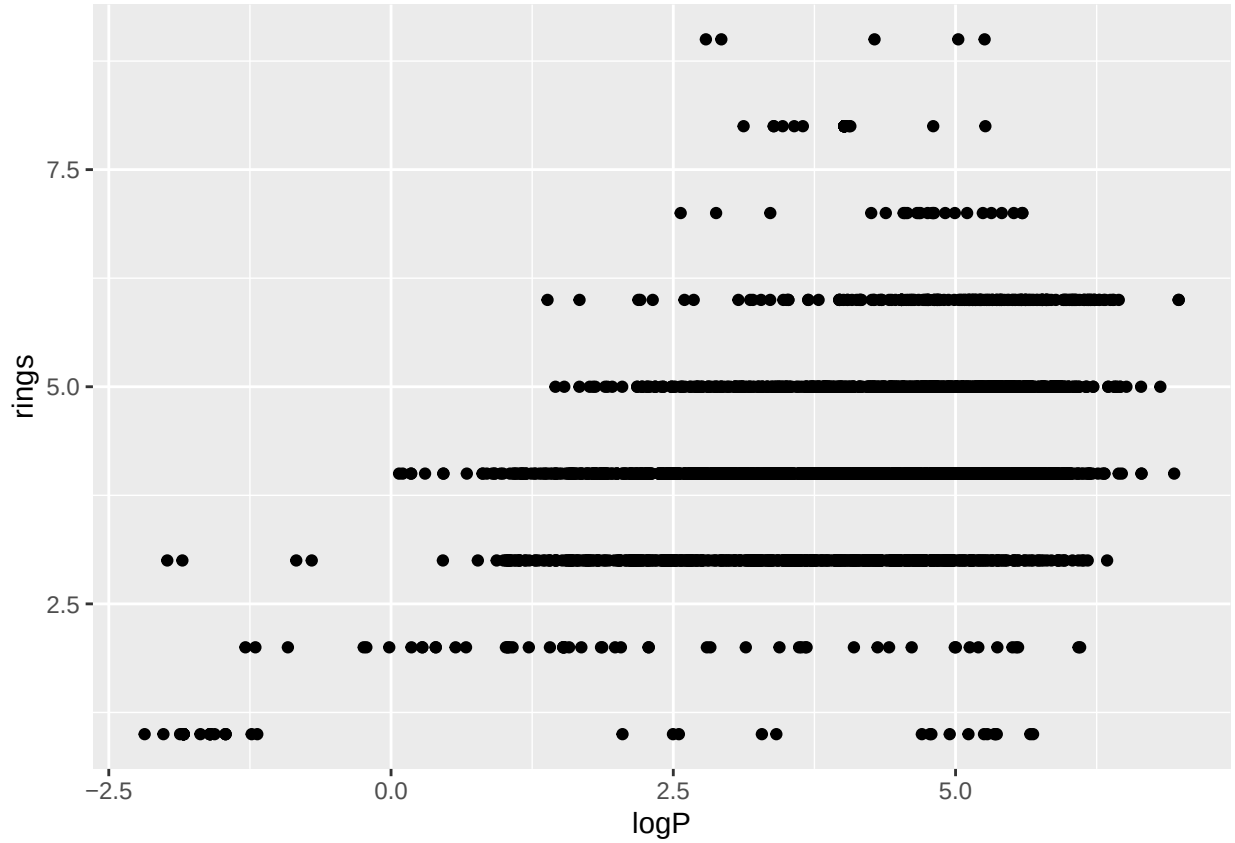


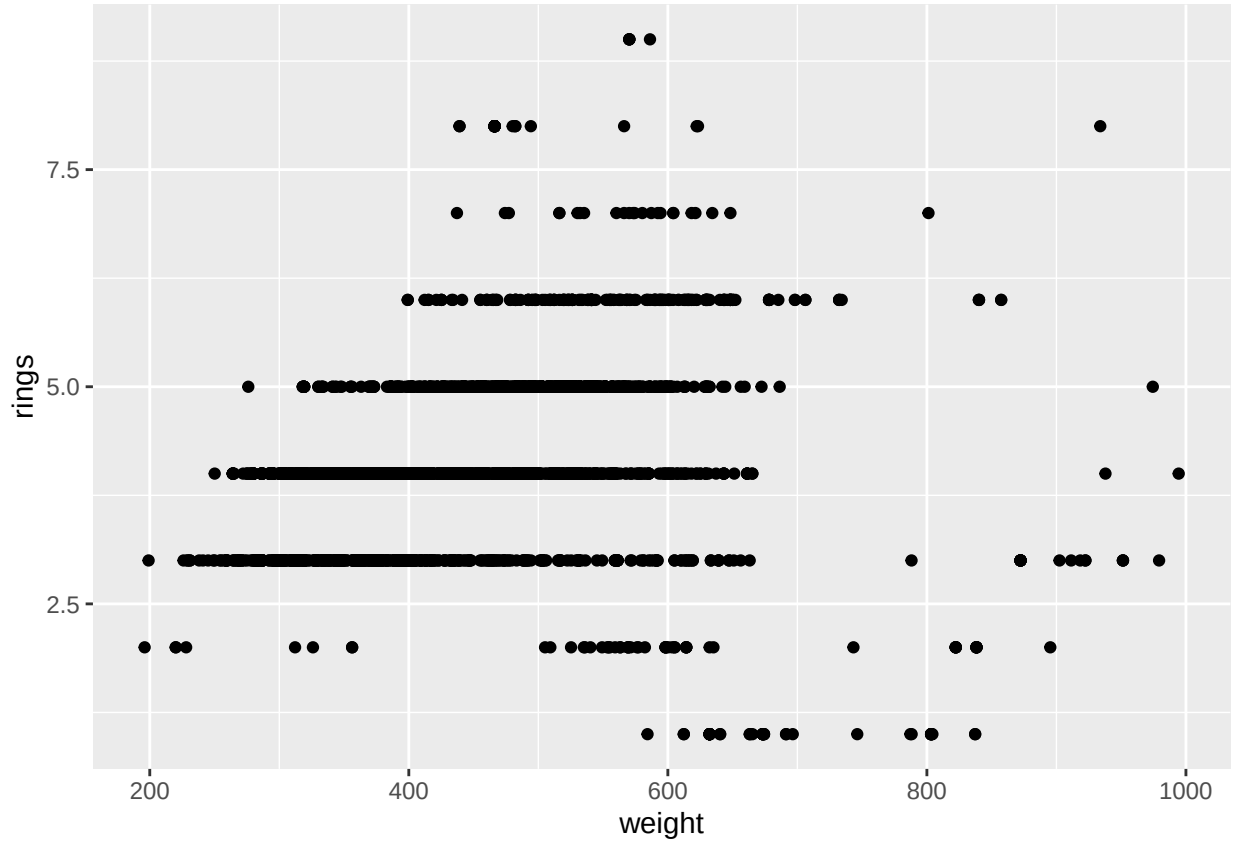
Scatterplot Visualization of Covariates to Predictor Variable

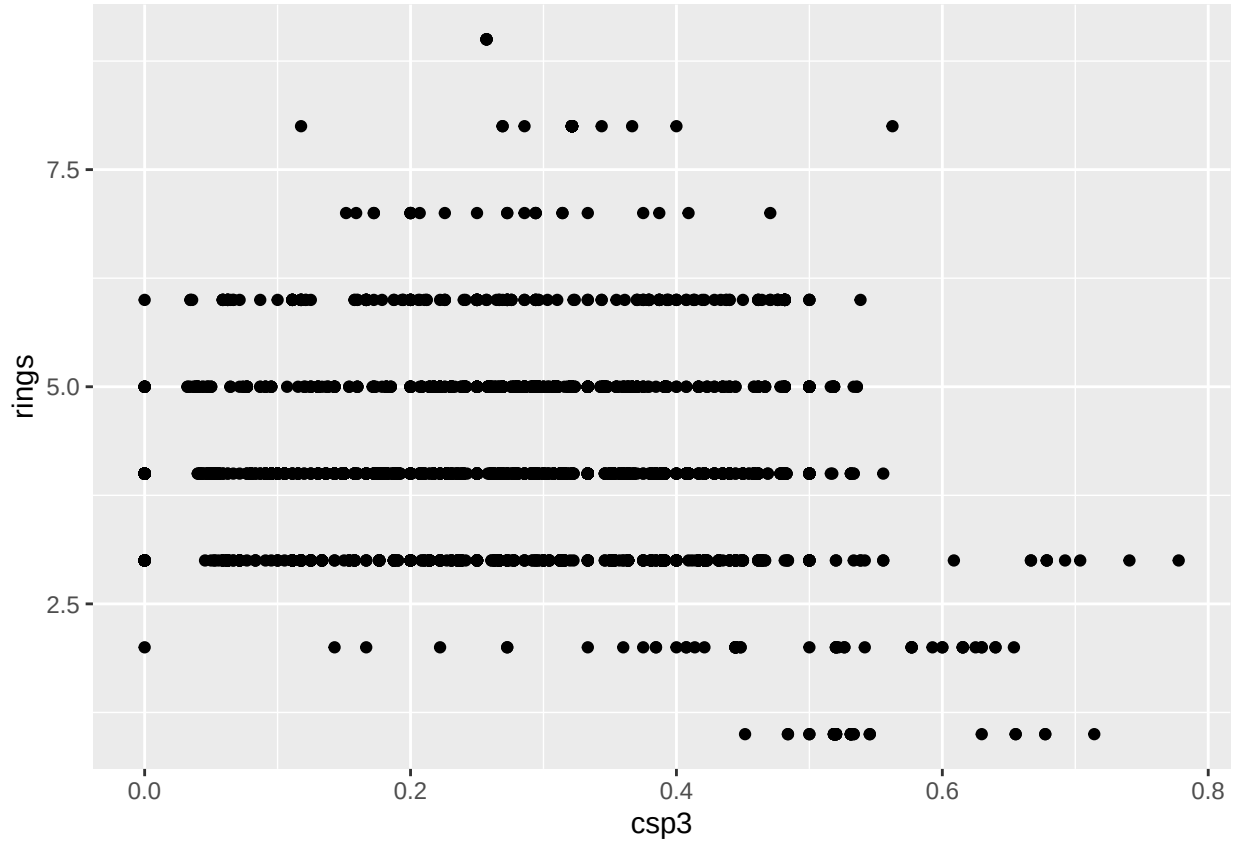
The scatter plot visualization helps in understanding how each variable is correlated to the predictor variable. In some ways we can understand what type and how big of a role each variable plays in the prediction and potentially define the correlation.

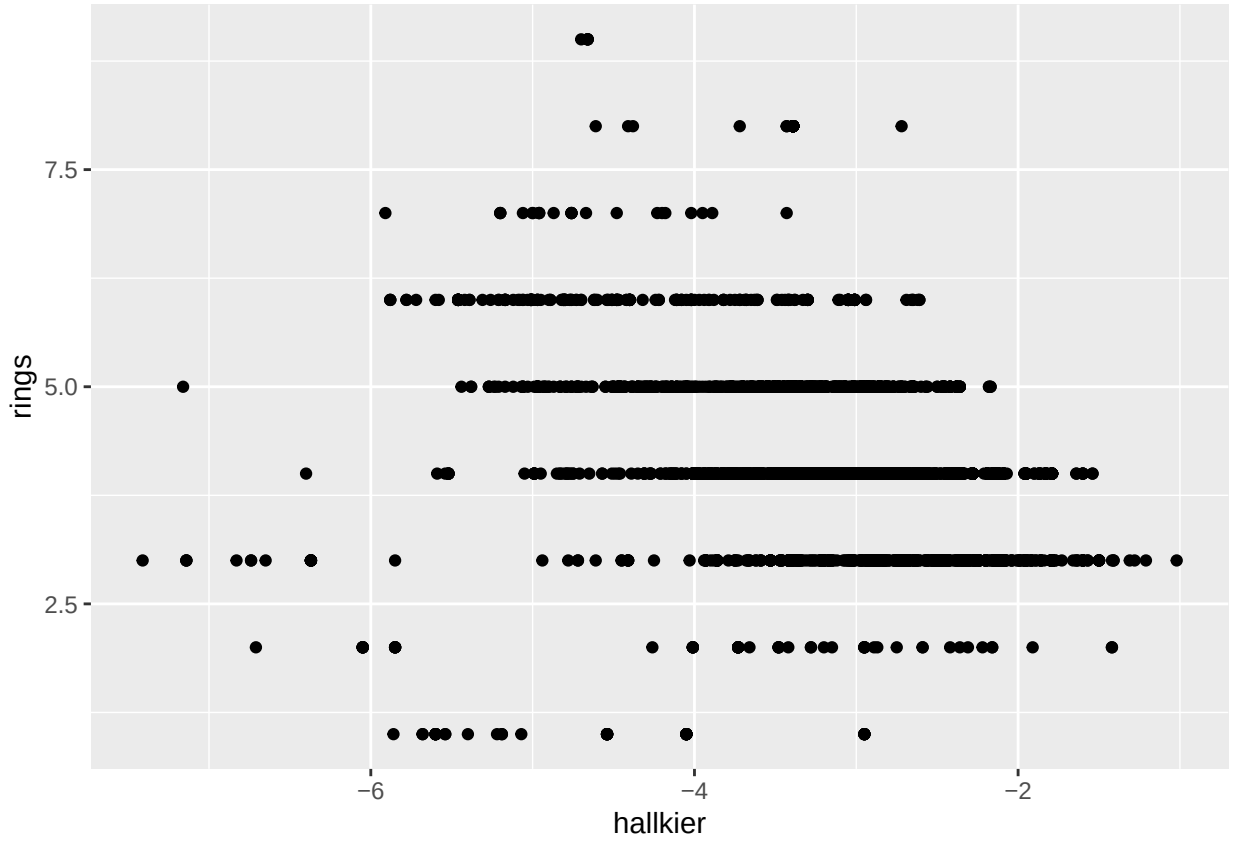

rings
7.5
5.0
2.5
0.4
0.5
0.6
0.7
0.8
qed

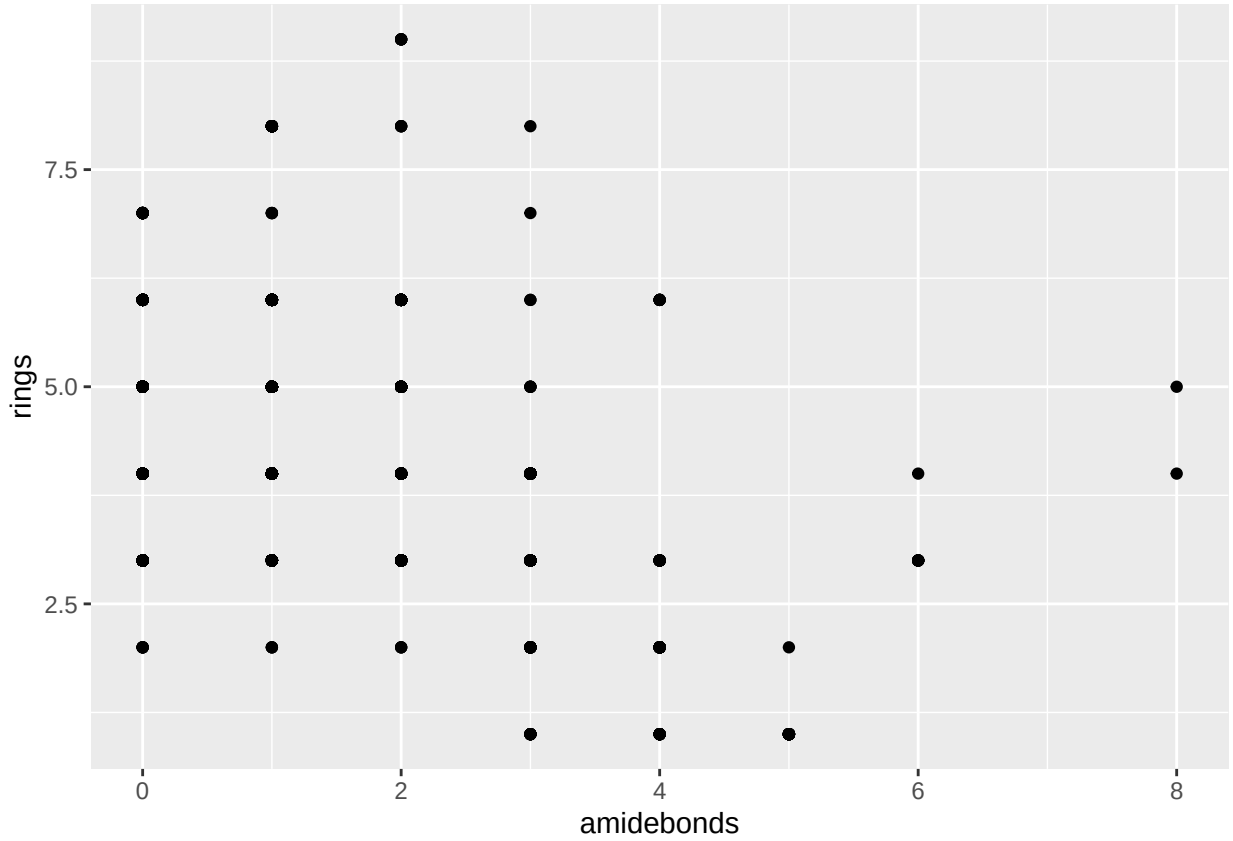

rings
7.5
5.0
2.5
2
3
4
5
SAS

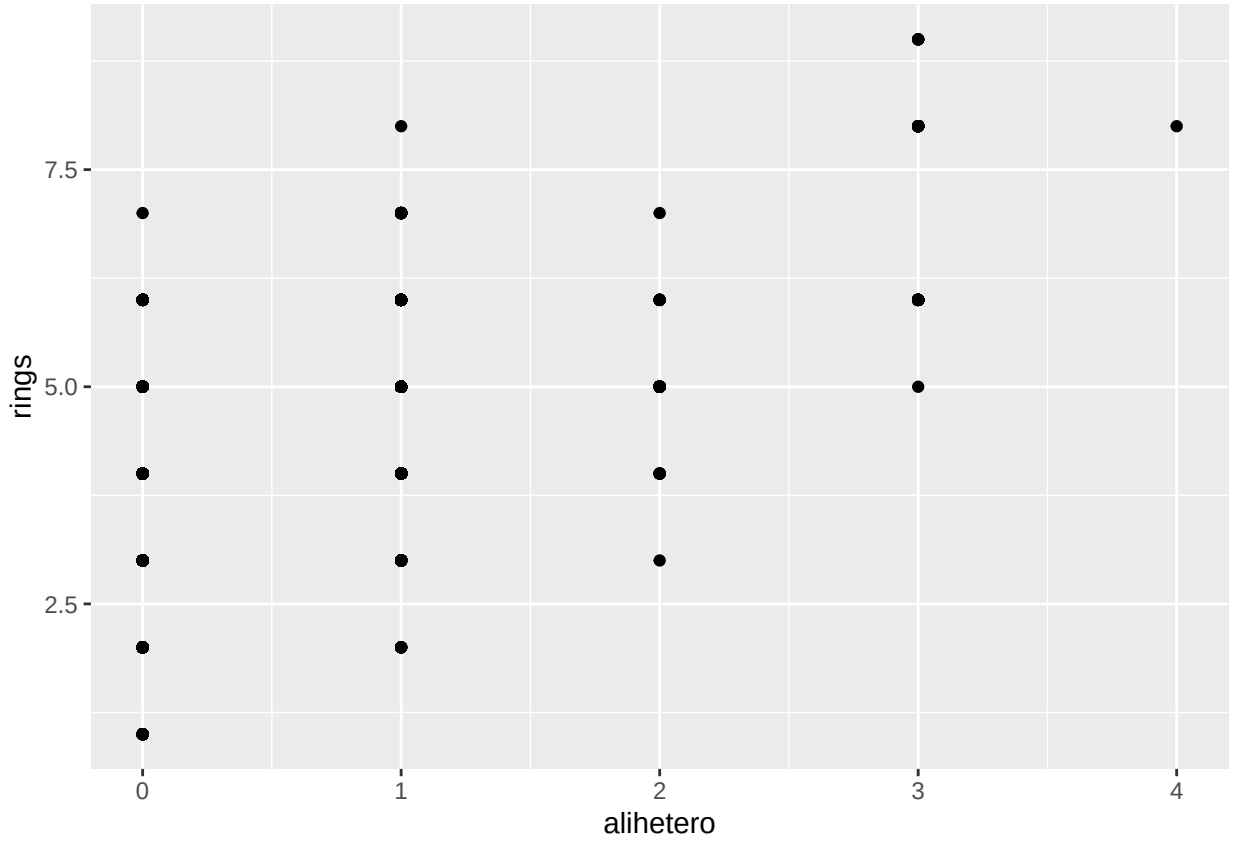

rings
7.5
5.0
2.5
−2.5
0.0
2.5
5.0
logP

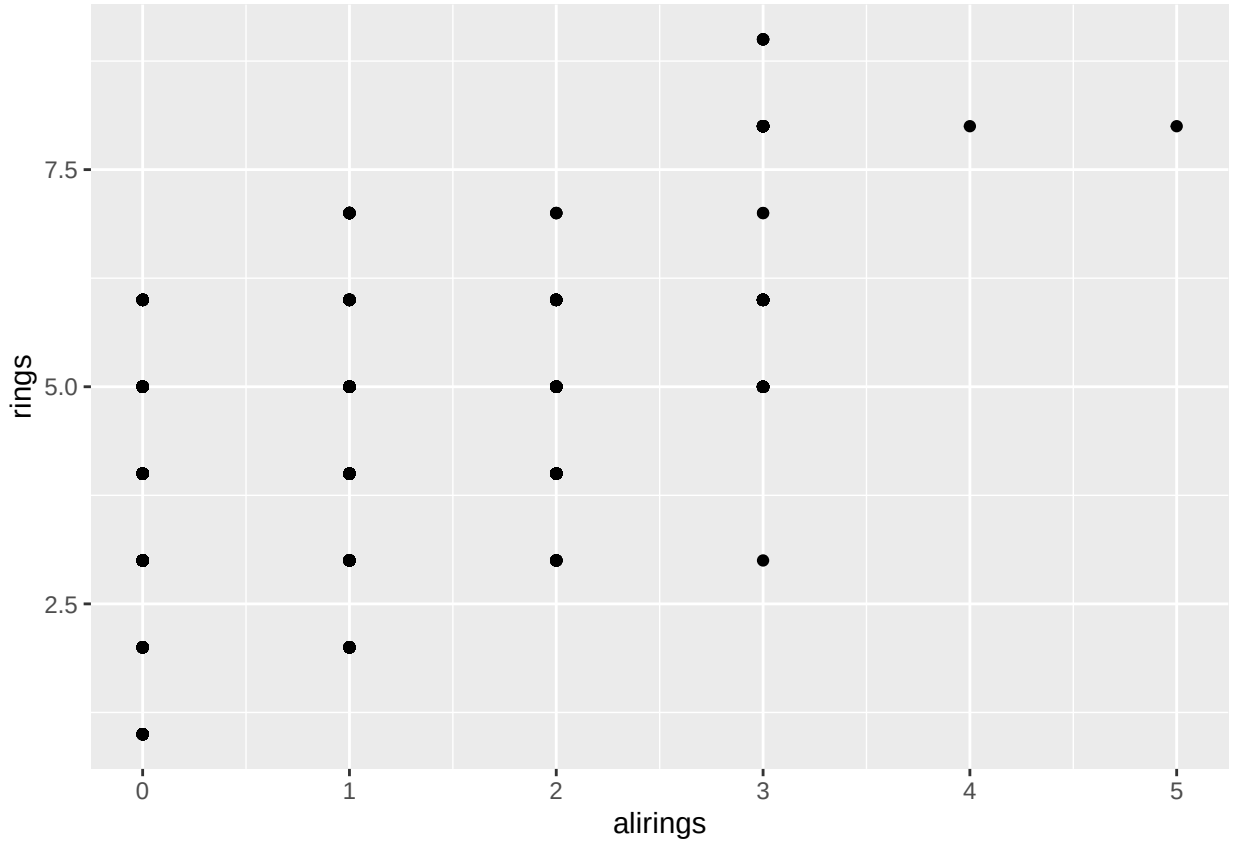

rings
7.5
5.0
2.5
200
400
600
800
1000
weight

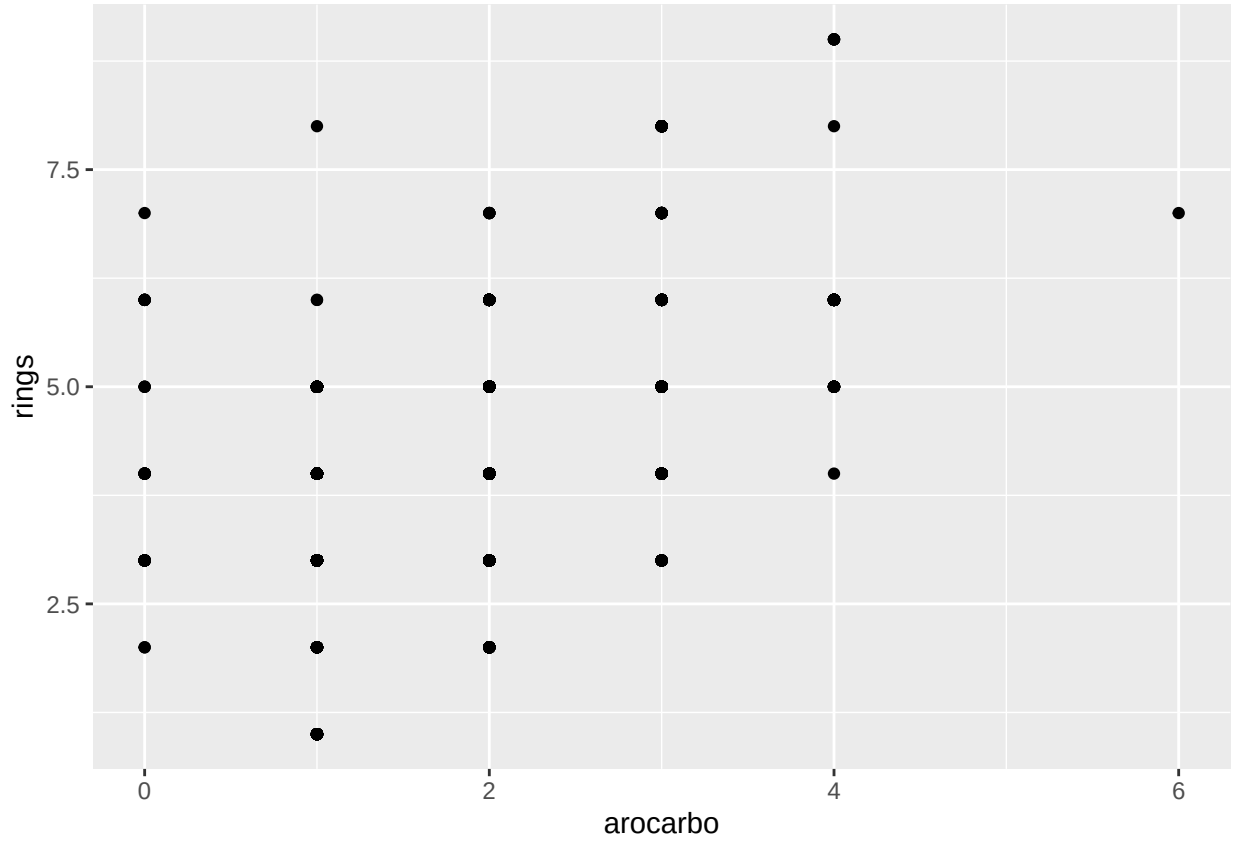

rings
7.5
5.0
2.5
0.0
0.2
0.4
0.6
0.8
csp3

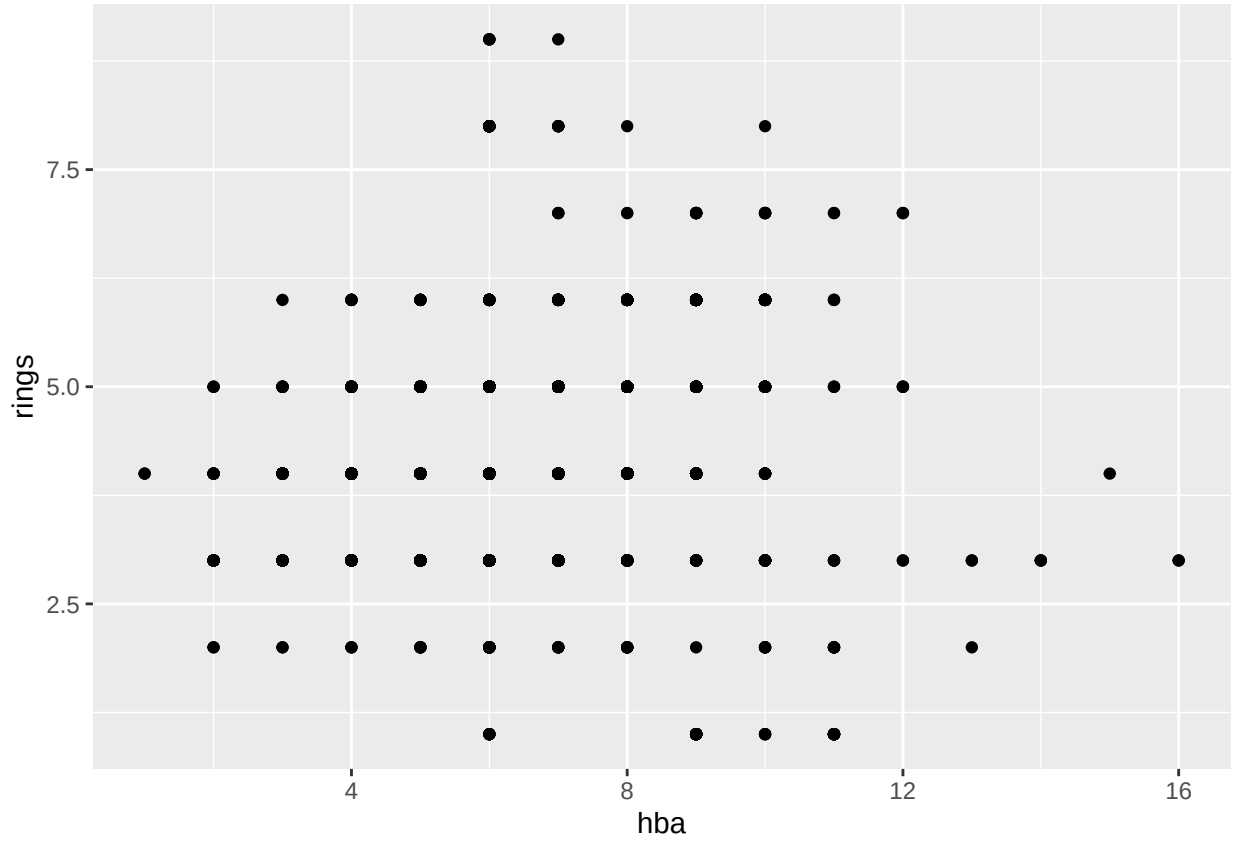

rings
7.5
5.0
2.5
−6
−4
−2
hallkier

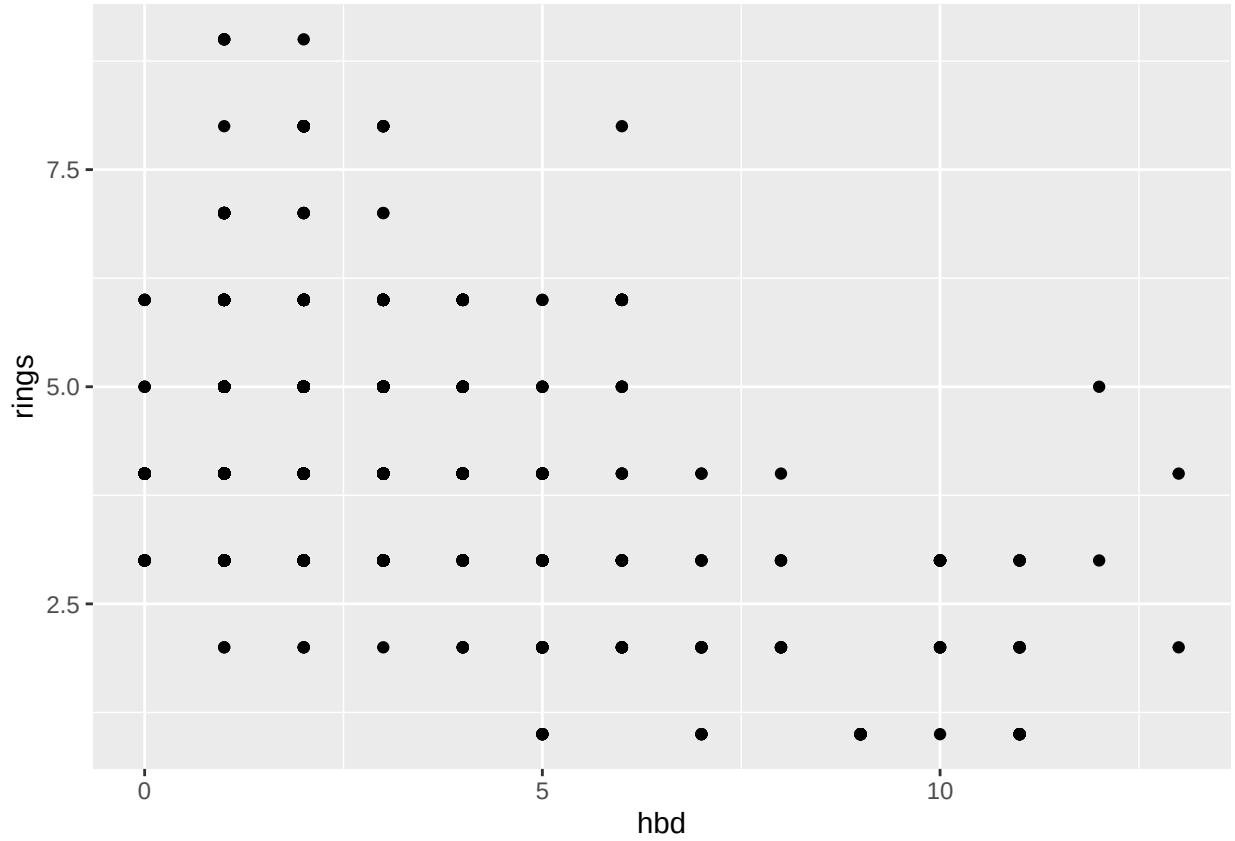

rings
7.5
5.0
2.5
0
2
4
6
8
amidebonds

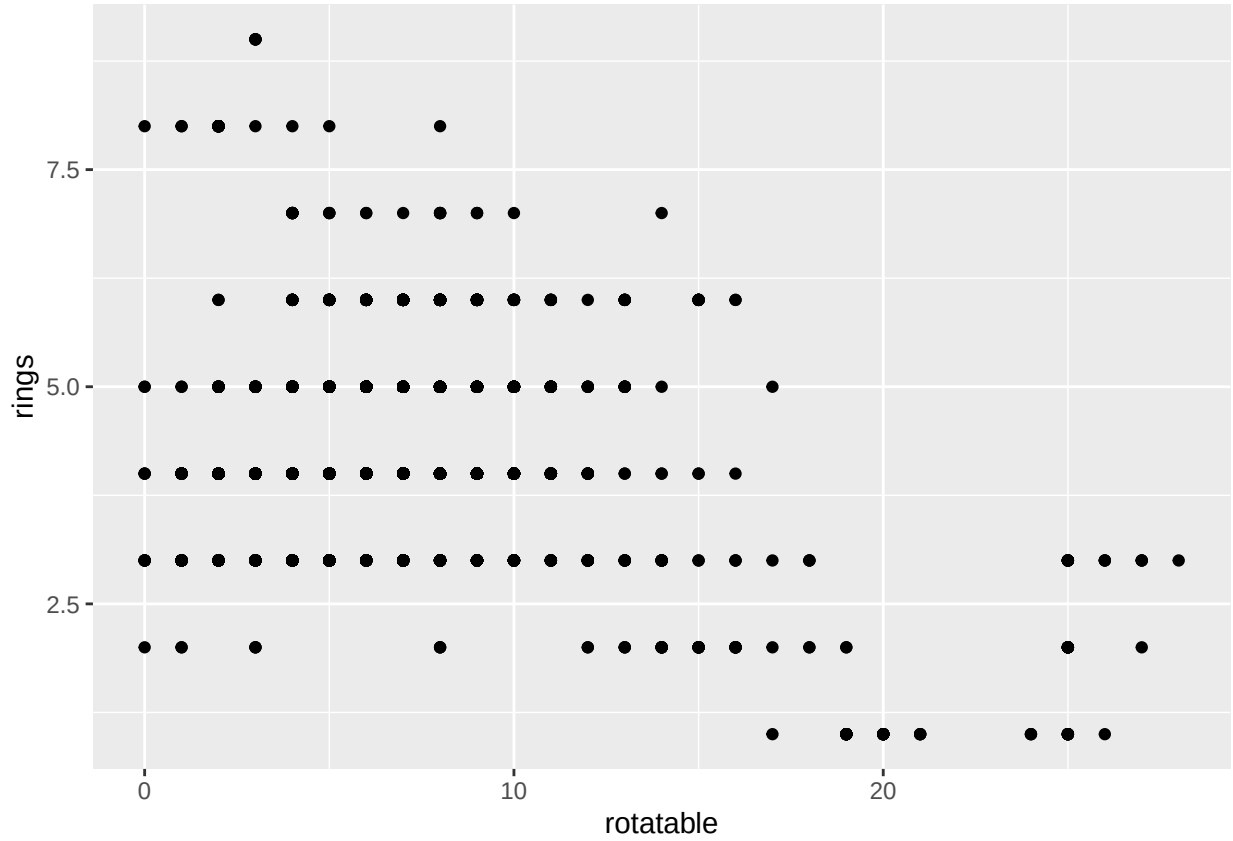

rings
7.5
5.0
2.5
0
1
2
3
4
alihetero

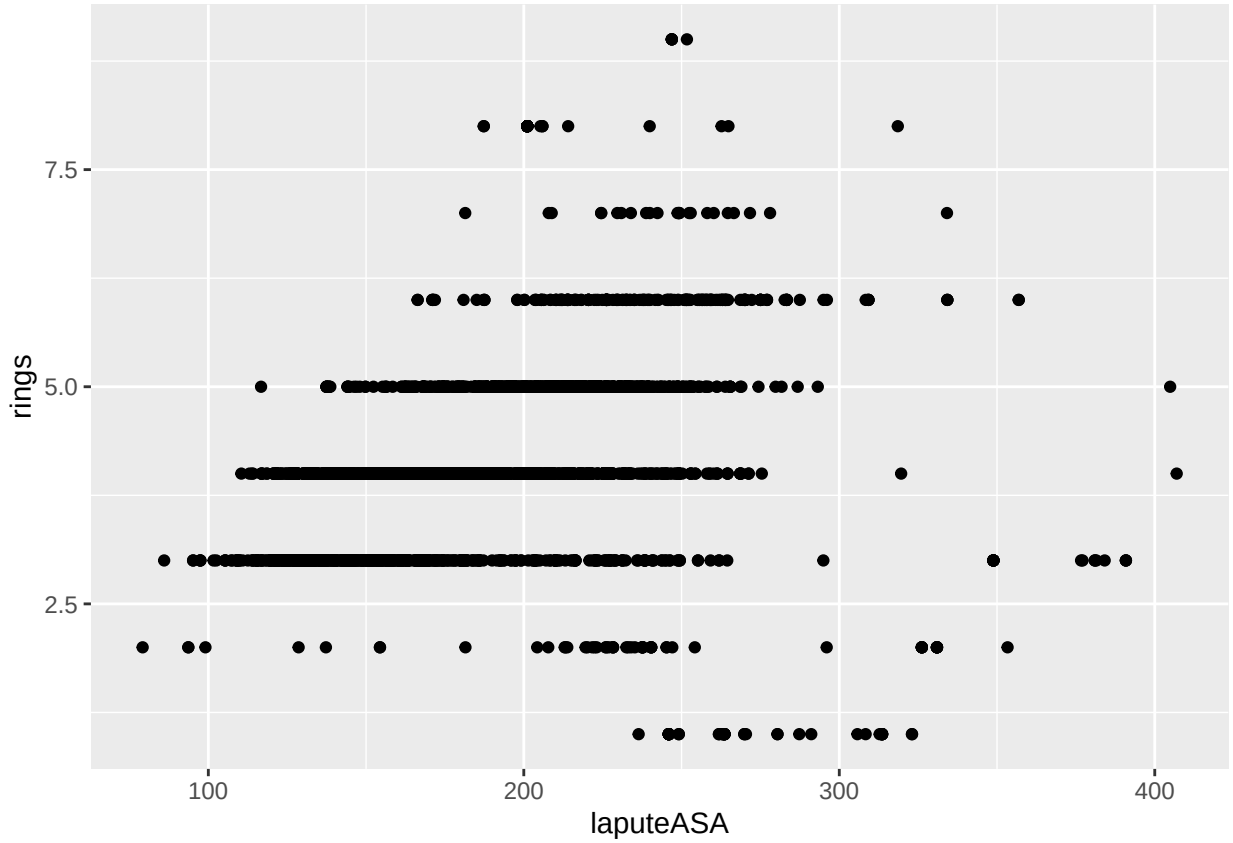

rings
7.5
5.0
2.5
0
1
2
3
4
5
alirings

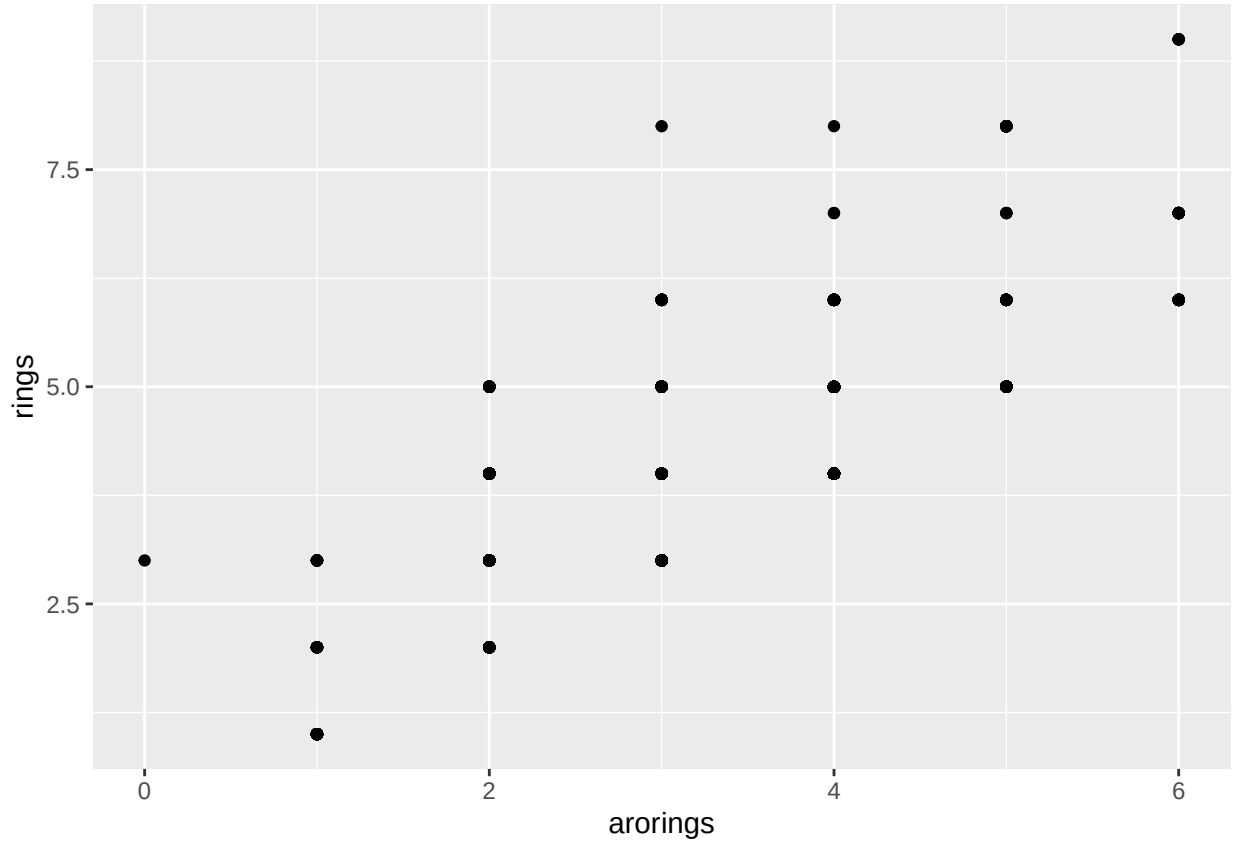

rings
7.5
5.0
2.5
0
2
4
6
arocarbo

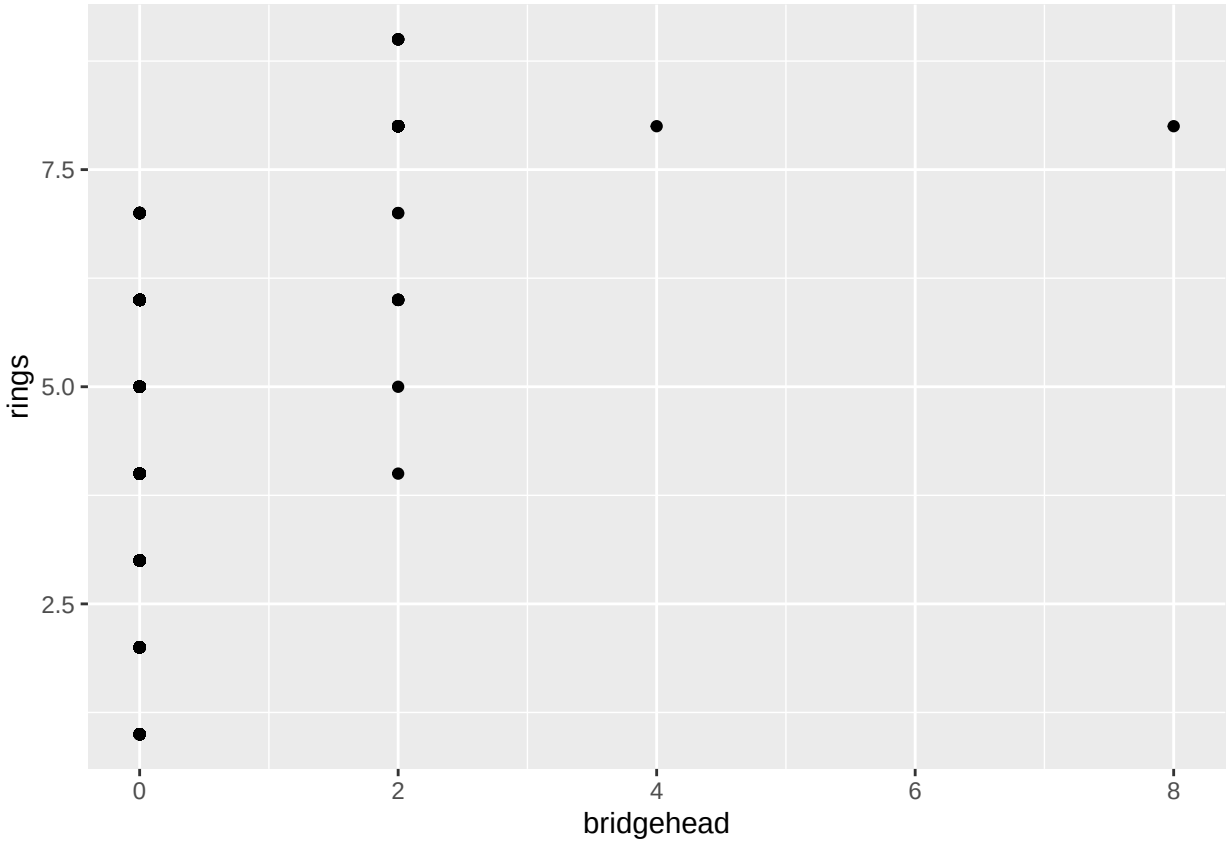

rings
7.5
5.0
2.5
4
8
12
16
hba


rings
7.5
5.0
2.5
0
5
10
hbd


rings
7.5
5.0
2.5
0
10
20
rotatable

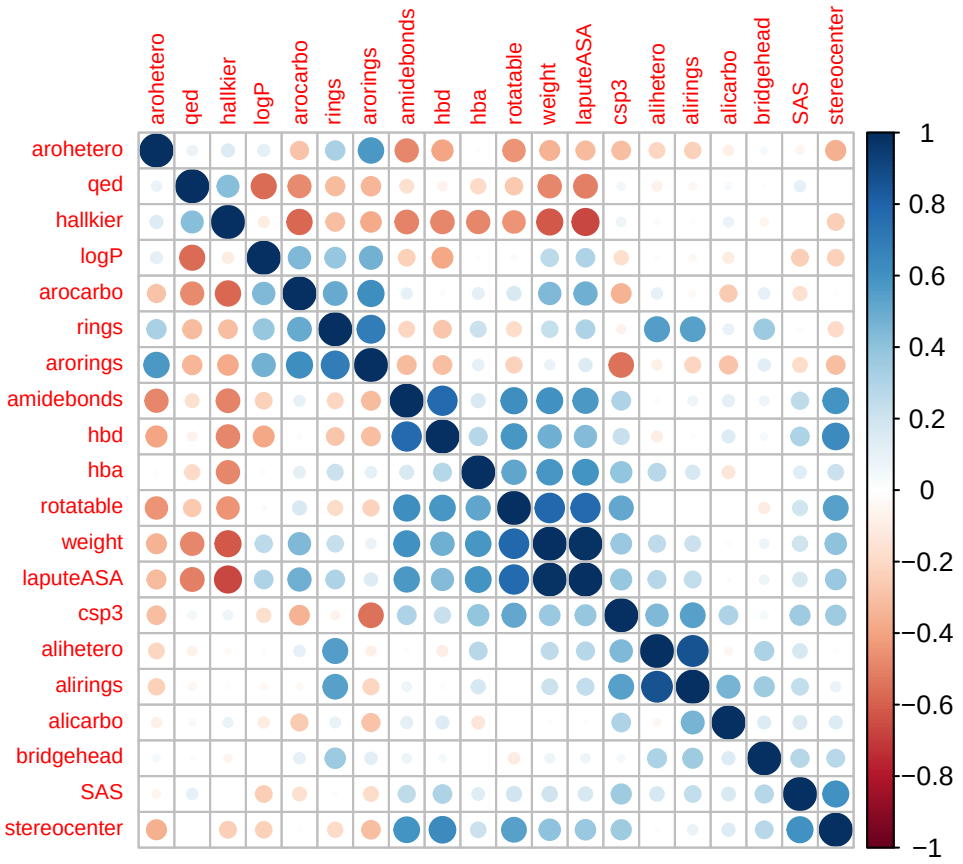

rings
7.5
5.0
2.5
100
200
300
400
laputeASA


rings
7.5
5.0
2.5
0
2
4
6
arorings



Correlation Plot of Covariates

The correlation plot gives us insight into how well the variables interact with each other and how co-linear or correlated each variable is to one another. This correlation plot is a visual representation of a correlation matrix for the feature set. The circles closer to dark blue indicate higher correlation and the circles closer to red indicate less correlation. This also aids us in understanding the data better, especially in which variables will have huge effect on the predictor variables.



Now that we have completed the pre-processing and initial analysis of the variables, we move into creating training and testing sets of our data for the predictive models. The outcome variable we would like to predict is the "Rings" variable and the feature set is based on the other chemical attributes calculated. The split of the data is 60% training and 40% testing

The training and testing data sets will be used for the model building and learning, mostly in the Random Forest and Neural Network Model.

Model Building and Analysis

Now I will do some model building. Here we will analyze the performance of 4 models and determine the best fit model for the data: Linear Regression Model, Log-Linear Regression Model, Random Forest Regression Model, Neural Network Model.

The predictor variable is the number of rings and we have around 20 covariates that will be in the model.

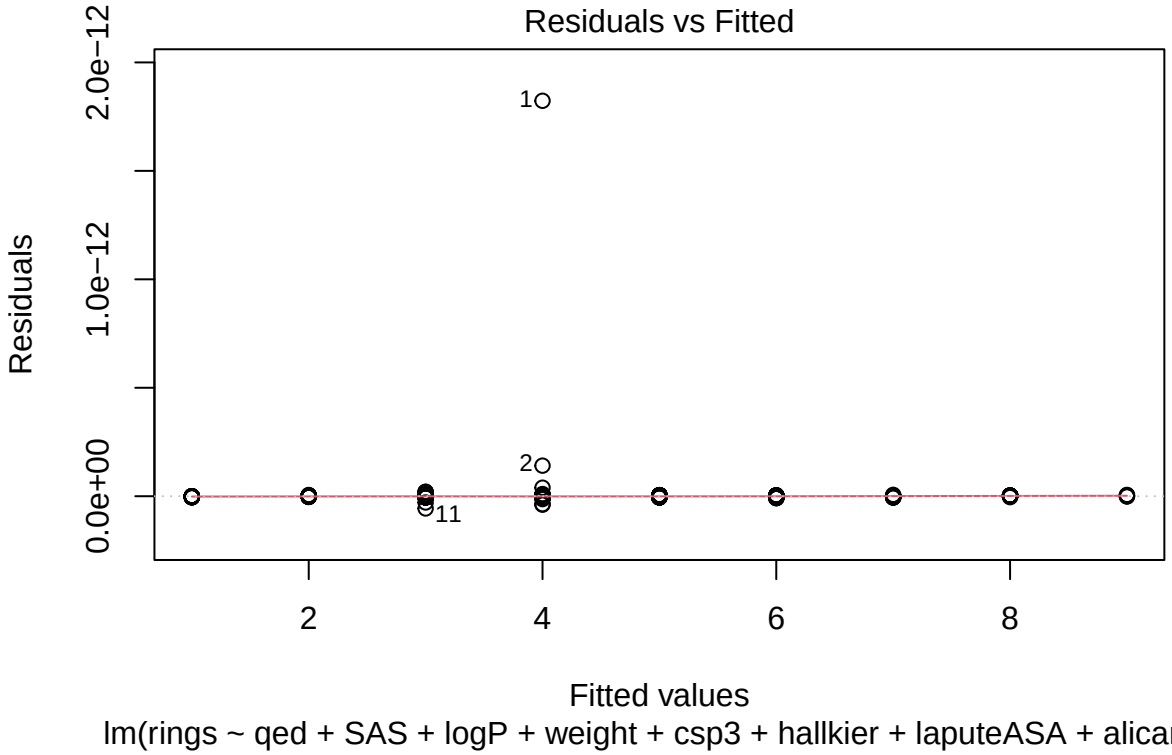
Multiple Linear Regression Model

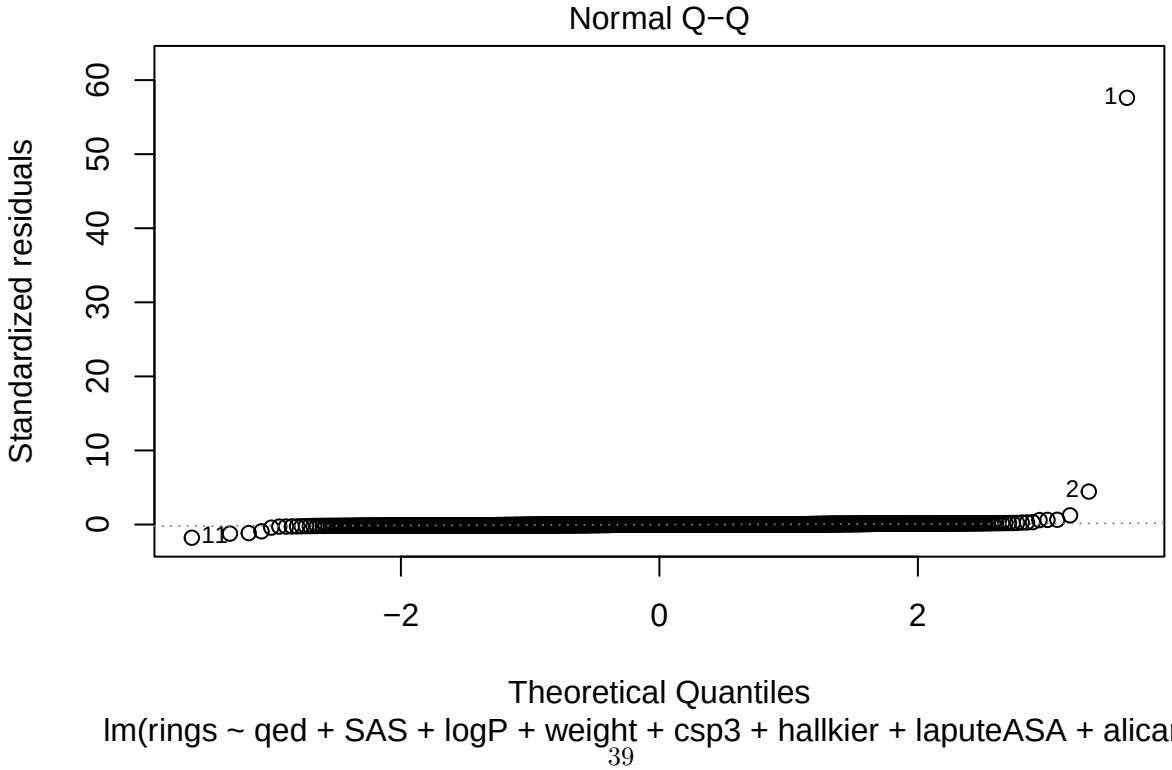
Multiple regression analysis (MR) is a highly flexible system for examining the relationship of a collection of independent variables to a single dependent variable. The independent variables may be quantitative or categorical.[10] The Multiple Linear Regression model is one of the basic but powerful predictive models that exist and they really help in assessing co-variates that have high impact on the predictor variable. The model is fitted such that each co-variate has its own effect size and impact on the predictor variable, using the Least Squares minimization of error method to come to an accurate fit. We evaluate the accuracy of the model using the residual diagnostics, where the residual refers to the difference between the actual values of the outcome and the predicted values, and the R2 value which explains the amount of variance from the

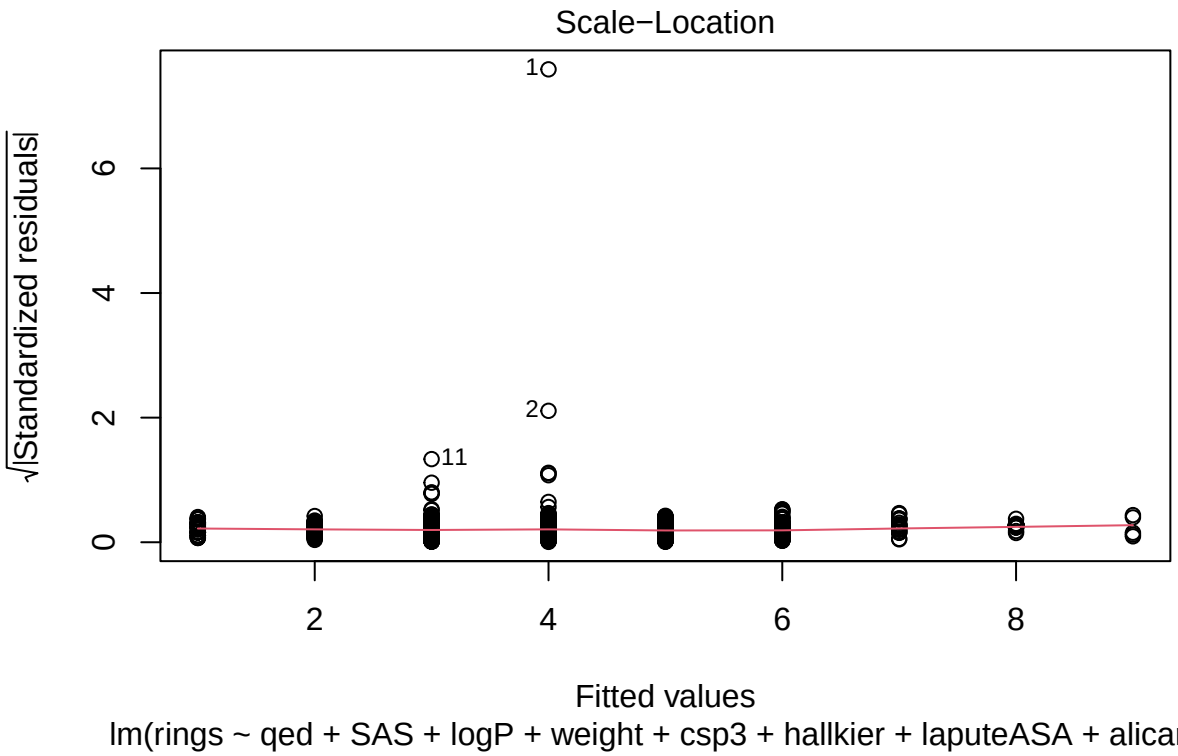
data that the model is accounting for. An ideal model has an R2 that is near the value of 1, thus accounting for most of the variance in the data set. Through our graphical analysis and interpretation, we will come to a conclusion on the effectiveness of this model for the data.

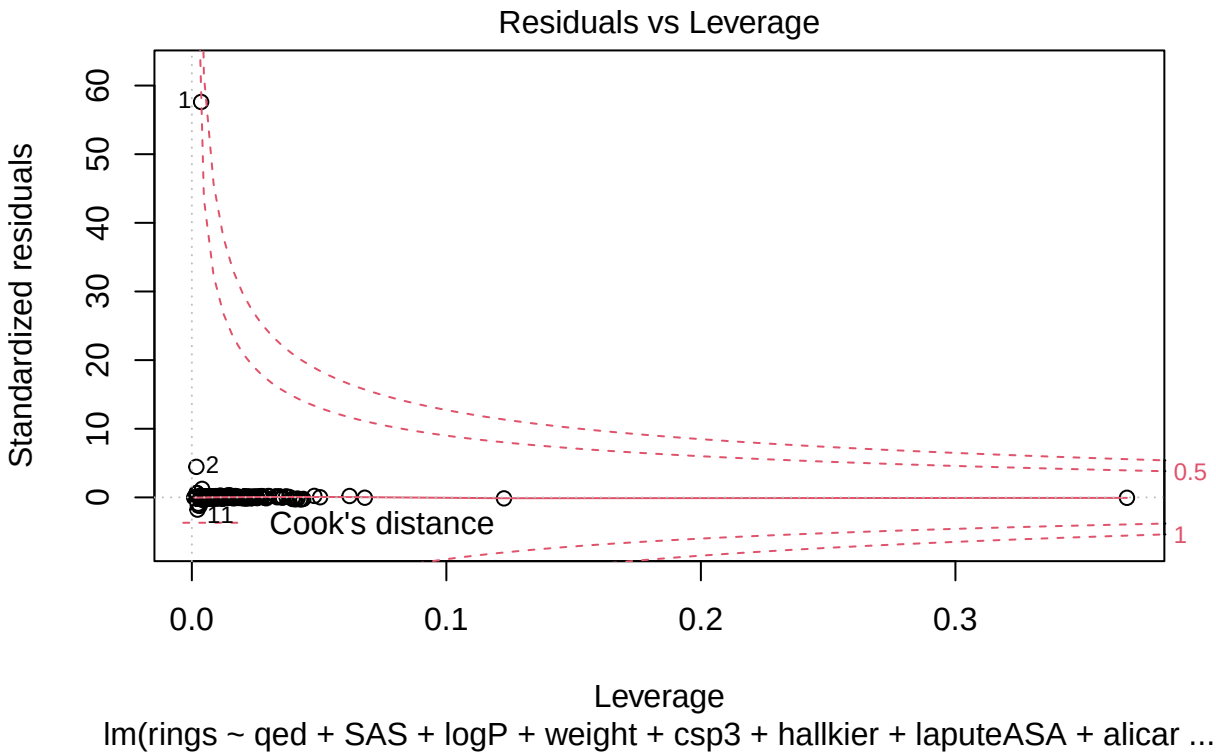
In this step, we start off by taking the rings as our outcome variable and regress it among all the co-variates in our data set.

Multiple Linear Regression Model 1






Scale−Location
√|Standardized residuals|
0
2
4
6
1
2
11
2
4
6
8
Fitted values
lm(rings ~ qed + SAS + logP + weight + csp3 + hallkier + laputeASA + alicar ...

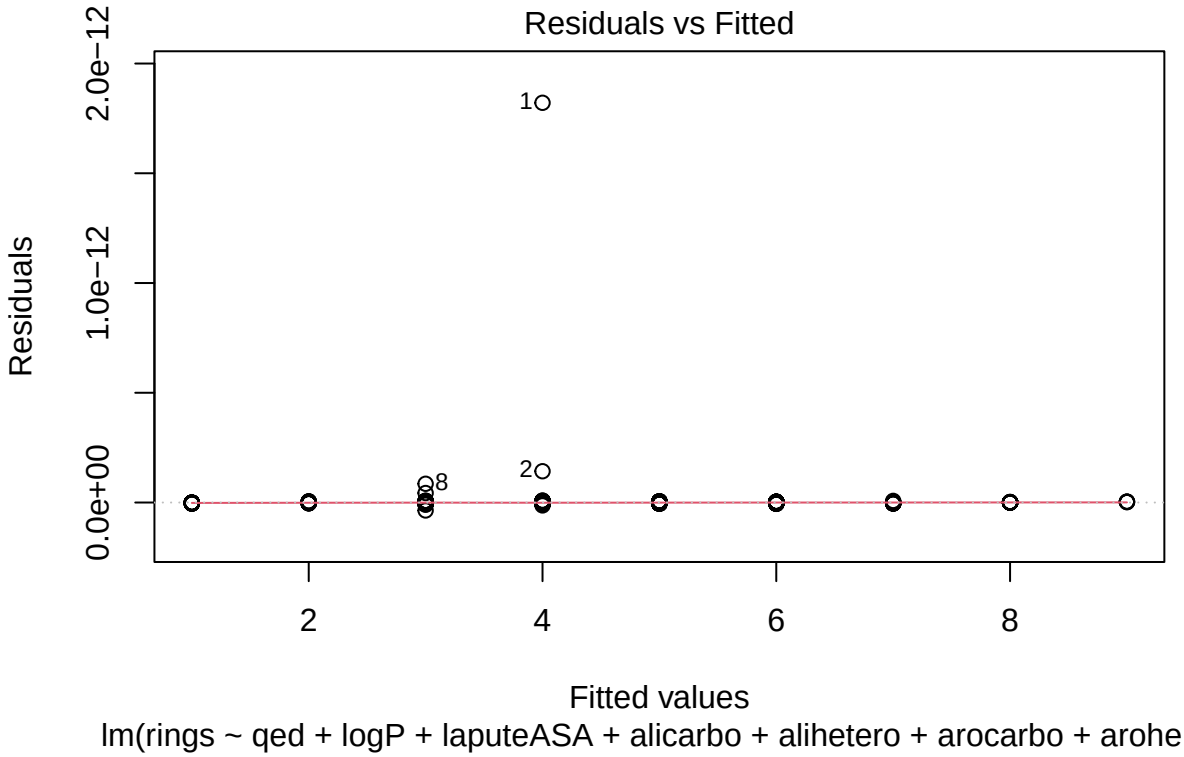


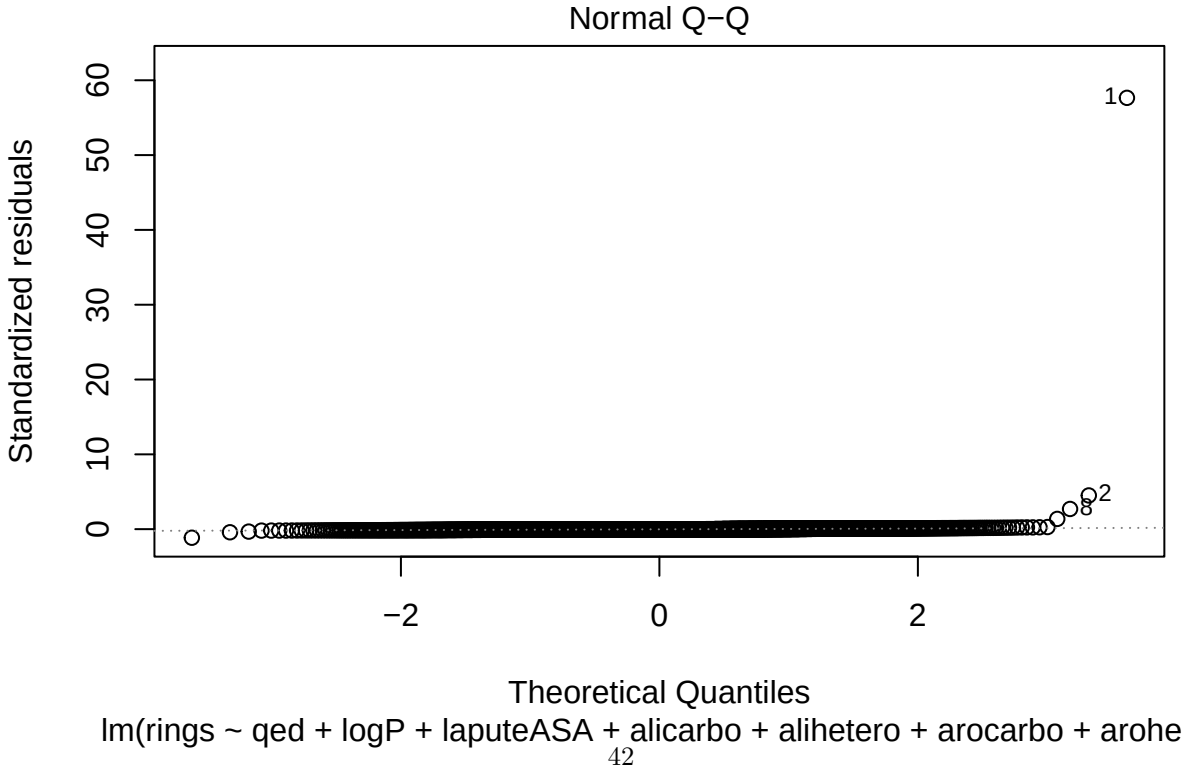
What we just achieved was building a basic linear regression model where we predicted the number of rings based on all of the co-variates we have in our data. By building the initial model, we get an R2 of 1, which plays cause for concern. The fact that the R2 is 1 tells us immediately that the model is over-fitting the data. If we also take a look at the graphs:

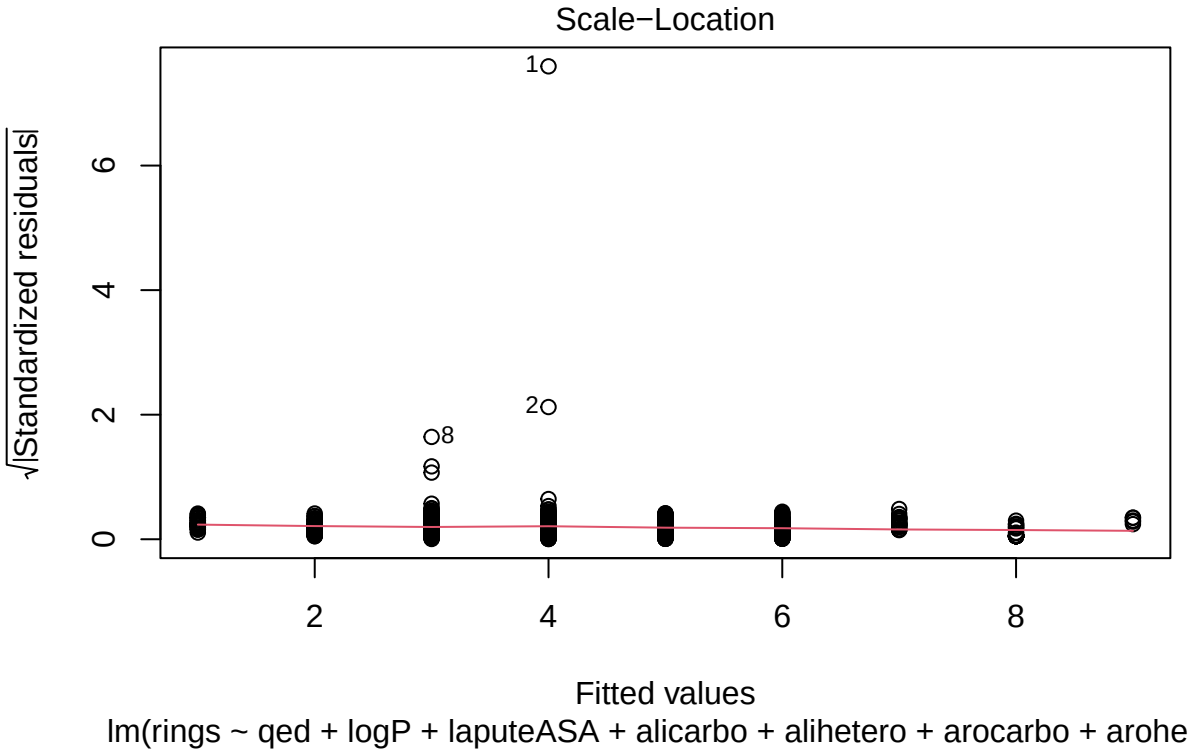
1. The residual versus fitted graph has no even distribution of points across the red line which tells us that the model is not really accounting for the variance even though it is saying it is. You can also see that the points are clustered together in locations, which further advocates for the stated.
2. In the QQ plot, there is really no curve of the plotted line which shows indication of overfitting.
3. We see both clusters of residual points in the 3rd graph as well as clustering in the Cook's Leverage graph. This can also be understood that, although our model tells us that it is accounting for 100% of the variance through our R2, we don't see any graphical indication of a good model. Additionally, the bounds for the Cook's Leverage graph are pretty low, indicating a bad fit.

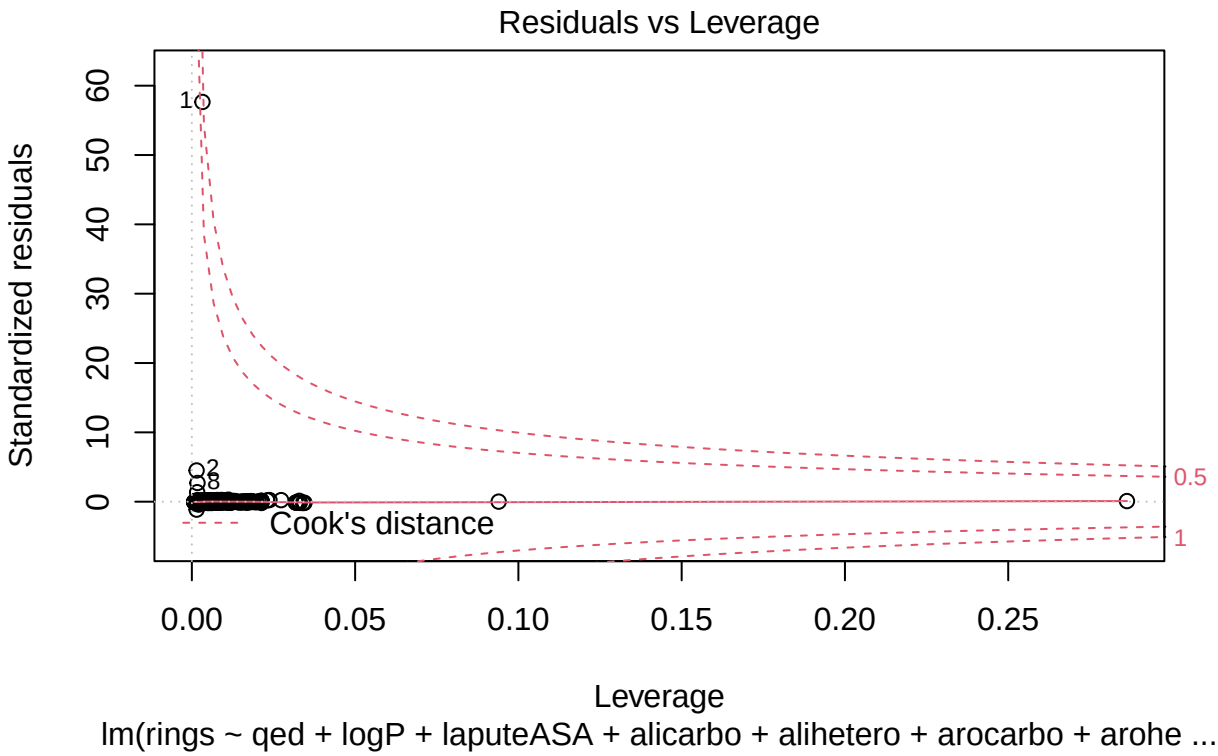
A key cause of this could be due to the amount of features in the model. There are significant features in the model such as alicarbo and alihetero, acting as main effects for the prediction, however, we see a good amount of features that are playing an insignificant role. Therefore, we start off by implementing Backwards Elimination of variables to rid the model of insignificant variables and re-fit our linear model based on the values that are significant.

Multiple Linear Regression Model 2






Scale−Location
√|Standardized residuals|
0
2
4
6
1
2
8
2
4
6
8
Fitted values
lm(rings ~ qed + logP + laputeASA + alicarbo + alihetero + arocarbo + arohe ...



As you can see, even after eliminating unnecessary features and fitting with the features we found somewhat or highly significant, we still end up with a skewed model where there is still clustering of residual points, over-fitting shown by the R2 value of 1 and graphical analysis. Therefore, we see that the Linear Regression Model is not the best fit model for the data and predictor variable, and we move on to the next model.

Log-Linear Regression

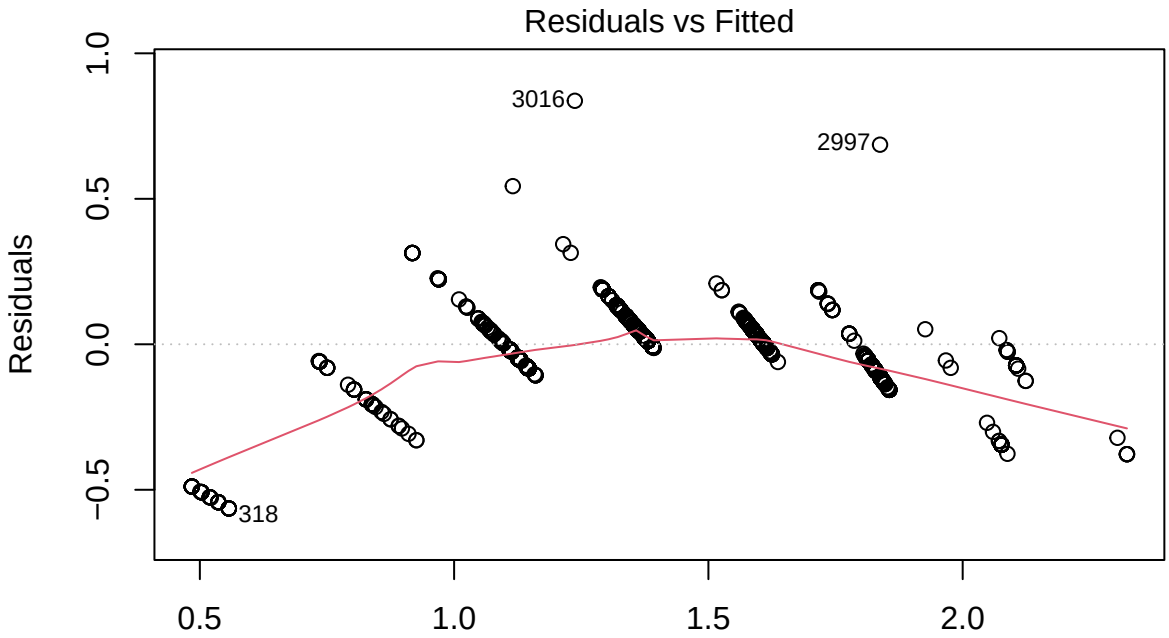
Next, we will attempt to use a basic log-linear regression model to see whether it will fit better compared to a linear regression model.

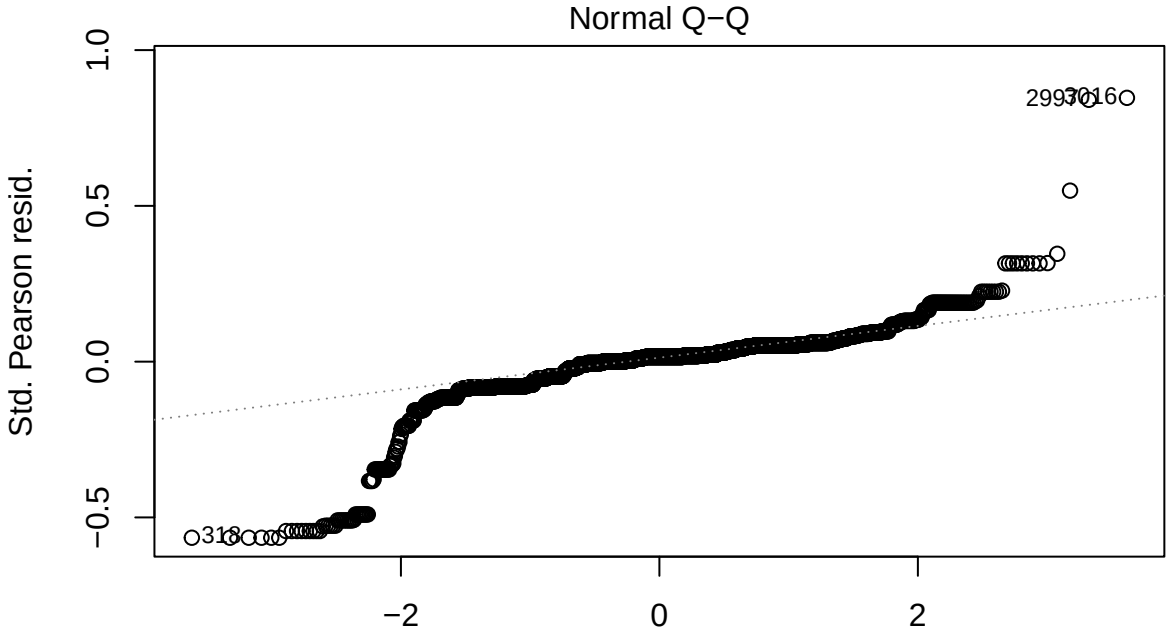
Log-Linear models are used to model predictor variables that act as count variables, such as number of hospital beds over a certain month, etc.

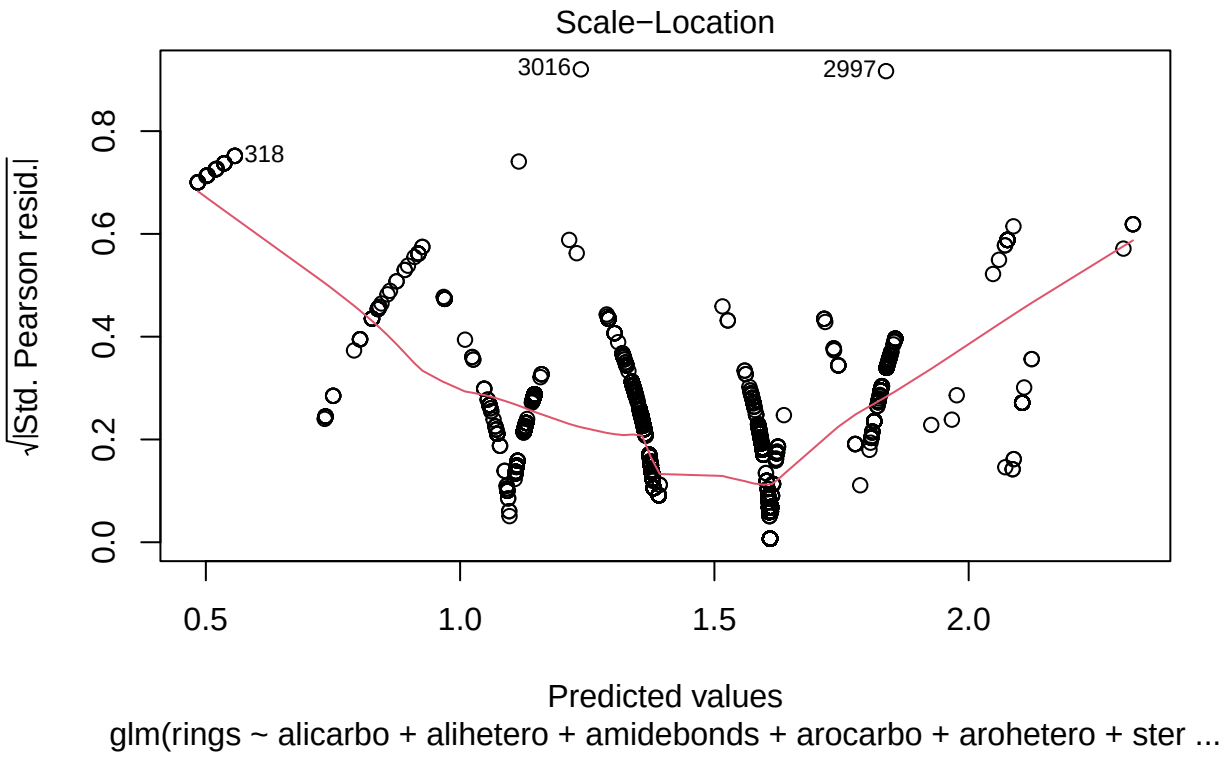
Log-Linear models are evaluated based on AIC and residual distributions as well. We also look into coefficients and significance of variables to draw conclusions about the model's effectiveness for prediction of our outcome variable.

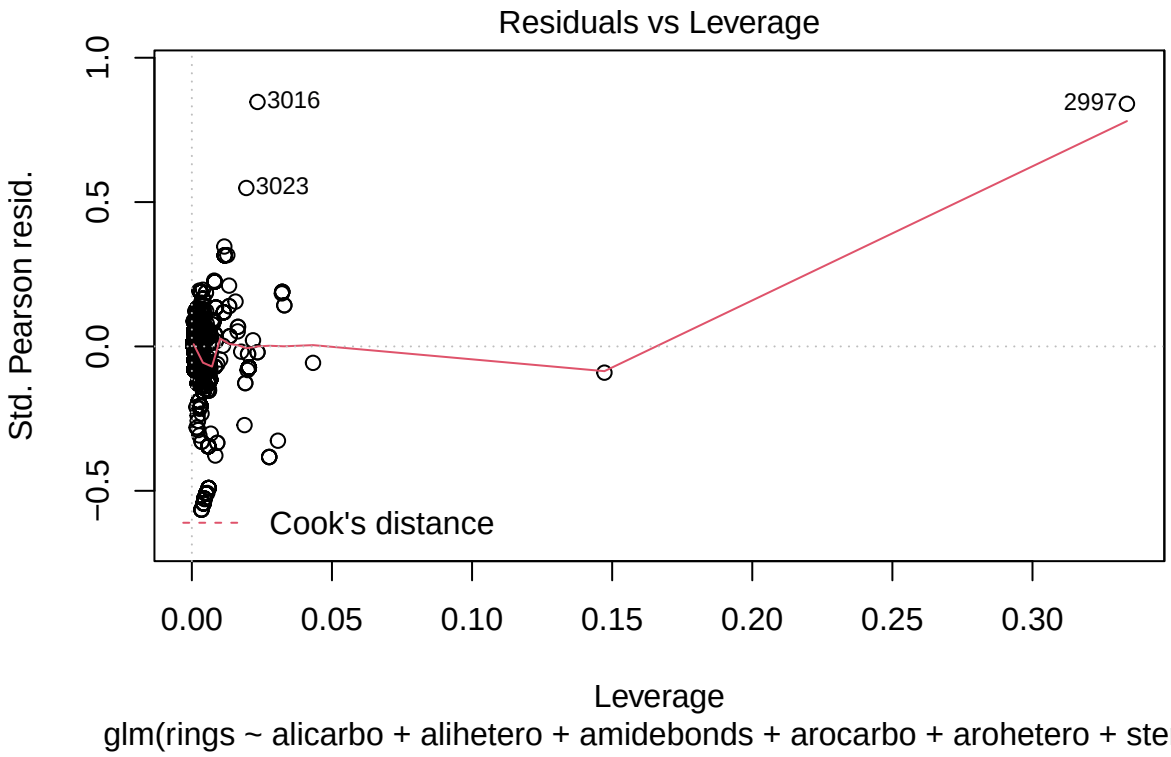
The reason we choose a log-linear model is due to the fact the number of aromatic rings(our predictor variable) can technically also be considered as a count variable. When we attempted to fit a linear regression onto the model, we ended up with a skewed and over-fitted representation of the predicted data, so let us try to interpret the outcome variable as a count variable to see whether the log-linear regression model will give us a better predictive model compared to linear regression.

Log-Linear Model






Scale−Location
3016
2997
318
√|Std. Pearson resid.|
0.0
0.2
0.4
0.6
0.8
0.5
1.0
1.5
2.0
Predicted values
glm(rings ~ alicarbo + alihetero + amidebonds + arocarbo + arohetero + ster ...



As we can see by the results of the log-linear regression model, we ultimately get a reasonable fit of the model to the data.

In this model too, we see that alicarbo, alihetero and some other co-variates that we saw in the last model prove to be very significant predictors and have important effects on the outcome variable.

The graphs show us that there is reasonable distribution of the residual points on the residuals versus fitted graph. We still see clustering of residual points however, which still indicates skewed model performance. We also see that there is even splittage of points below and above the red line of the residuals plot which portrays that this model is fitting somewhat reasonably. The QQ plot shows some skewage and deviation due to certain outliers of the data. The Cook's Leverage plot also shows us some outliers which could be skewing the data and the model fit. When we look at the AIC however, it is very high, indicating poor performance of the model. Although interpreting the predictor variable as a count variable was helpful, it still did not lead us into finding a good enough model that fits the data well.

As shown above, the log-linear model does a better job of fitting the data and predicting the outcome variable, however, the model's performance is pretty horrible given the AIC value and the distributions of the residuals through the graphical analysis.

The linear and log-linear models could also not fit due to the fact that the data has too much variance and is complex. With around 3000 instances and 20 features, we can expect to see this and it can tell us that a basic predictive model is not going to do well on this data. Therefore, we must use more complex algorithms that can account for the variance and complexity of the data. We move onto using a Random Forest model and analyze its accuracy.

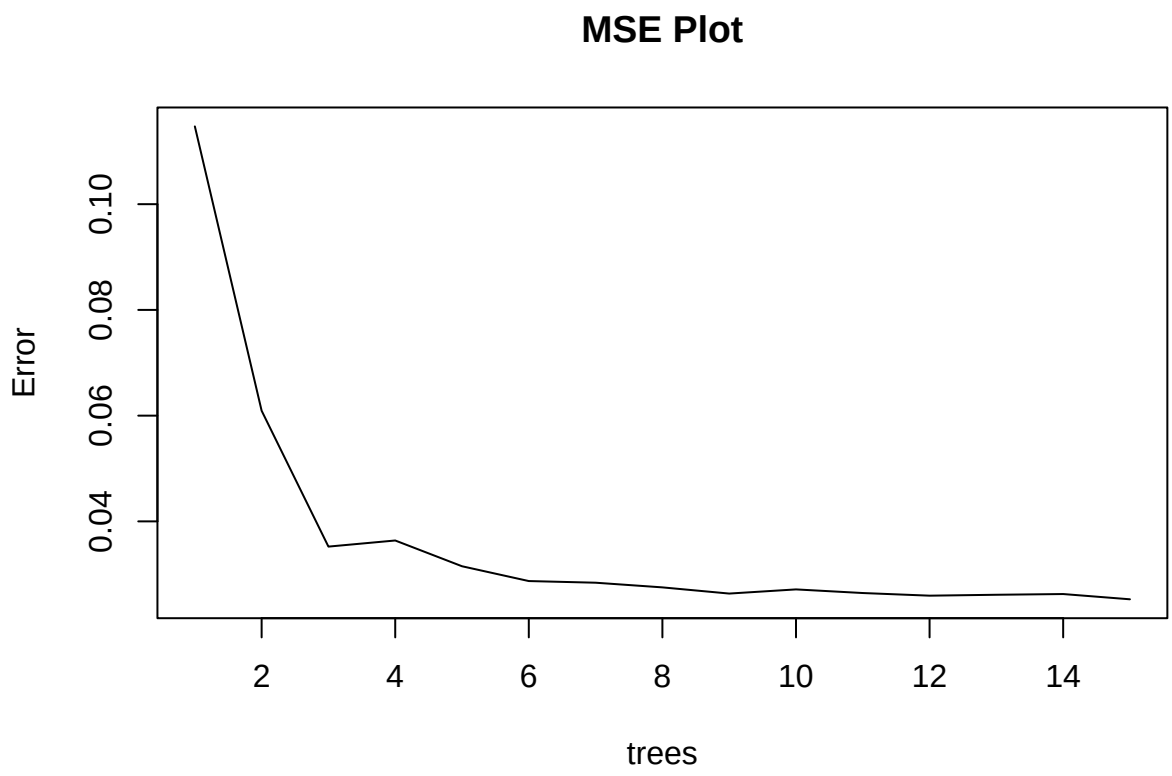
Random Forest Model

Now we move into higher techniques as we have determined that our predictor variable is more complex than we thought and we might need to use complex models to account for the variance in the data.

Random Forest Models are tree based models that help in understanding the complexity and variance of the data for prediction. Random forests seek to effect such correlation reduction by a further injection of randomness. Instead of determining the optimal split of a given node of a(constituent) tree by evaluating all allowable splits on all co-variates, as is done with single tree methods or bagging, a subset of the co-variates drawn at random, is employed. Breiman argues that random forests (a) enjoy exceptional prediction accuracy, and (b) that this accuracy is attained for a wide range of settings of the single tuning parameter employed. In the next section we further detail the formulation of random forests, and reveal a potential role for a second tuning parameter.[11]

They are evaluated based on the MSE, the RMSE, and the accounted Variance for both training and testing sets. The tuning parameter is set as the number of trees and can be adjusted as you build your model. We attempt to fit a Random Forest model with certain parameters and see how well it is able to predict the ring count.

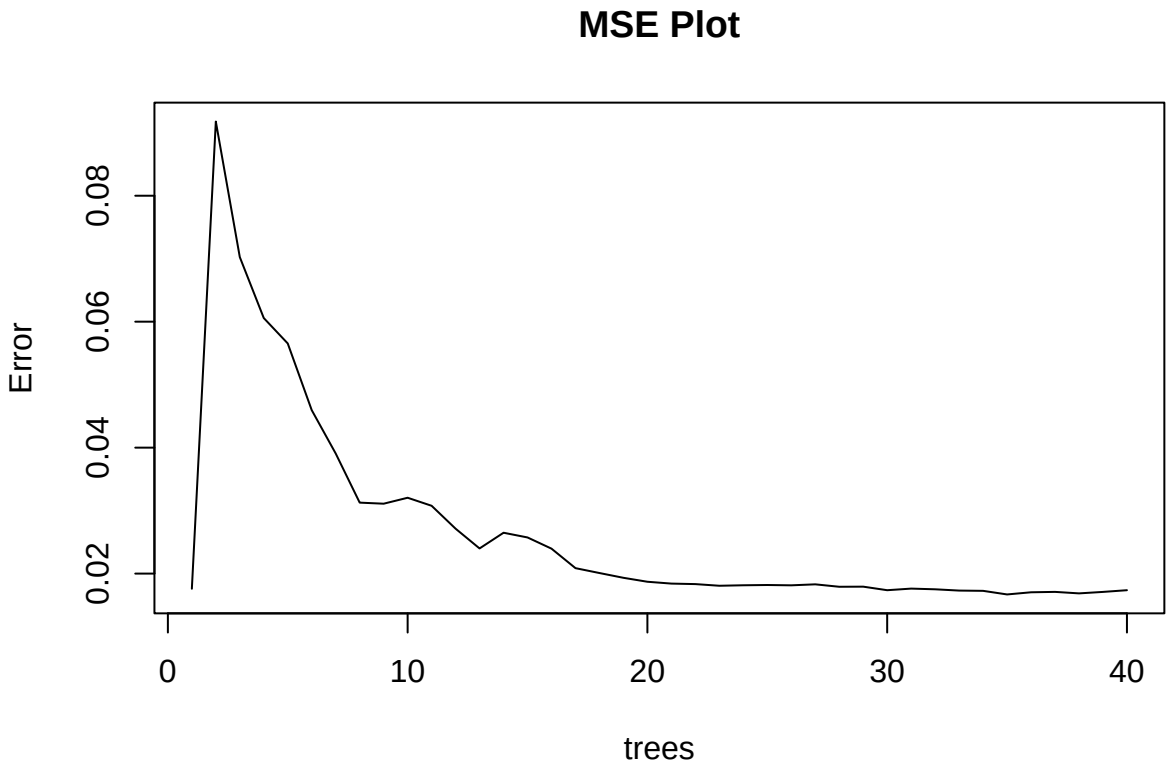
Random Forest Model 1



We fit a Random Forest model to the data with an initial parameter value of 15 trees. Through the output of our results, the Random Forest predicts very well with the mean of squared residuals of training set being around 0.015 and the testing set around 0.22. The percentage of variance accounted for being around 98%. Also looking at the MSE graph, we see that the error significantly decreases as the number of trees increase. This gives us insight into how well the model predicts given the data and ultimately tells us that this model is ideal to predict the number of aromatic rings.

To see if we can improve the model a little more, we increase our tree count to 40 trees and see whether we can reduce the mean of squared residuals value.

Random Forest Model 2



As said before, the increase of the number of trees does improve the model slightly as it brings the mean of squared residuals value down to 0.011 and the test set MSE is 0.02. The % of Variance explained in training is 98.89% and the variance explained for testing is 98.44% indicating good performance. The MSE graph also shows that the error decreases as the number of trees increase, so it is able to account for the variance in the data. All in all, the Random Forest model seems to be the best predictive model for the number of rings.

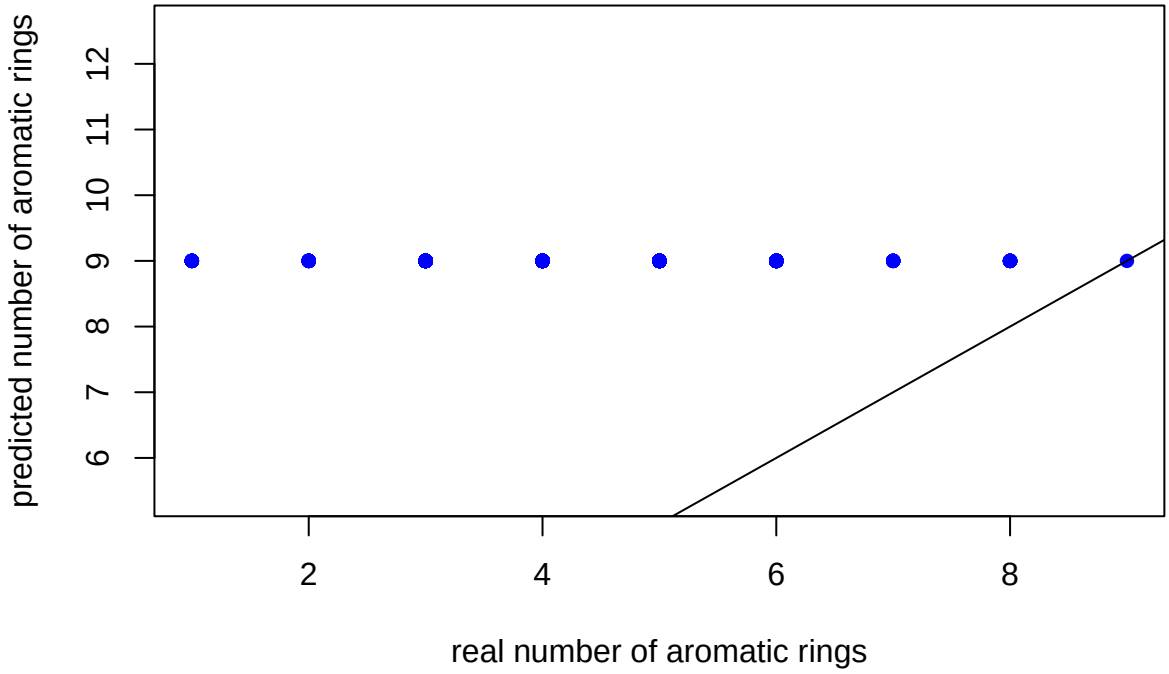
Now we try to implement a Neural Network model, just to see whether adding more complexity will cause the data to be misrepresented.

Neural Network Model

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process.[12]

The evaluation of the model will be done through RMSE statistic, the error, as well as the predicted vs actual values graphical analysis. There is an off chance that this model will be too complex to fit the data.

We attempt to fit a Neural Network model to the data, where it is a dense 3 layer network. The neural network model might or might not be a good fit for this data as the variable we are predicting is continuous and ultimately, it might be better to use one of the models from before.



As you can see, the neural network model is not good for the prediction of rings based on the data. The error is very high at around 1000 and the RMSE is around 30.04, higher than in the Random Forest, showing us that over-complication of model fitting can lead to misrepresenting the data. Since this problem is a regression based problem, using Neural Networks did not help with its predictive ability.

Conclusions and Further Remarks

SRC Kinase Inhibitors are critical molecules in the space of cancer and act as potential drug inhibitors that can be used for treatment today. Although there are other characteristics that play influential roles in determining the efficacy of these molecules, there are variables that play crucial part and are often overlooked. The confluence of both biological systems and computational techniques help to open up the door for more discovery and research. As this field is so vast, even the smallest contribution can add another piece to the puzzle, continuously improving our research in the area of cancer drug therapy.

In terms of the project, we achieved both of the objectives: the first one being to create a good model that will be an accurate model to predict the number of rings in an SRC Kinase Inhibitor and the second one being to do a comparative analysis of the different models we created and see which performs the best. Out of all the models built, I come to the conclusion that the Random Forest model with 40 trees performs the best on the data compared to all other models. It accounts for the complexity of the data as well as the variance, causing it to perform well in both training and testing. The MSE plots also give us insight into how well the error decreases as the number of trees increase, showing us that the model performs very well on this dataset. I think the main takeaway from this project would be to really understand how your data

is distributed and the variance it contains, utimately using that information to formulate a good predictive model. With the confluence of all the work put in, from initial pre-processing and visual analysis, to model building with numerical and graphical analysis, we ultimately achieved both our goals, leaving us satisfied.

Further steps I would take in this project would be to include a detailed PCA analysis for feature importance as well as being able to figure out a way to include features such as molecular formula or structural characteristics into this dataset. By implementing thes techniques, I think there might be a change in how the data represents the predicted variable and can lead to some new discovery.

References and Citations

1. Zhu, Y.; Alqahtani, S.; Hu, X. Aromatic Rings as Molecular Determinants for the Molecular Recognition of Protein Kinase Inhibitors. Molecules 2021, 26, 1776. https://doi.org/10.3390/molecules26061776

2. Manning, G.; Plowman, G.D.; Hunter, T.; Sudarsanam, S. Evolution of protein kinase signaling from yeast to man. Trends Biochem. Sci. 2002, 27, 514–520. [Google Scholar] [CrossRef]

3. Johnson, G.L.; Lapadat, R. Mitogen-activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinases. Science 2002, 298, 1911–1912.

4. Al-Obeidi, F.A.; Wu, J.J.; Lam, K.S. Protein tyrosine kinases: Structure, substrate specificity, and drug discovery. Biopolymers 1998, 47, 197–223.

5. Cohen, P. Protein kinases—The major drug targets of the twenty-first century? Nat. Rev. Drug Discov. 2002, 1, 309.

6. Hunter, T. Why nature chose phosphate to modify proteins. Philos. Trans. R. Soc. B Biol. Sci. 2012, 367, 2513.

7. Manning, G.; Whyte, D.B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. Science 2002, 298, 1912–1934

8. Hanks, S.K.; Quinn, A.M.; Hunter, T. The protein kinase family: Conserved features and deduced phylogeny of the catalytic domains. Science 1988,

9. Bridges, A.J. Chemical inhibitors of protein kinases. Chem. Rev. 2001, 101, 2541–2571.

10. Schneider A, Hommel G, Blettner M. Linear regression analysis: part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2010;107(44):776-782. doi:10.3238/arztebl.2010.0776

11. Segal, M. R. (2004). Machine Learning Benchmarks and Random Forest Regression. UCSF: Center for Bioinformatics and Molecular Biostatistics. Retrieved from https://escholarship.org/uc/item/35x3v9t4

12. Maind, S. B., & Wankar, P. (2014). Research paper on basic of artificial neural network. International Journal on Recent and Innovation Trends in Computing and Communication, 2(1), 96-100.