Structural Discovery of Macromolecules Data Analysis

Sam Kazan

2023-02-13

1 Structural Discovery of Macromolecules

The main question in structural discovery field is how macro molecules behave in their natural forms and natural environments and how macromolecules stray away from natural behavior.

2 Setup for Data Analysis

2.1 Loading Libraries and the Data Frame

I selected following libraries for this data analysis. They cover various aspects of data analysis. I use "ggplot2" for data visualization to uncover patterns and relationships in my data. "Dplyr" helps me manipulate data by filtering, summarizing, and transforming it. "Plotly" provides interactive visualizations and "dbscan" can identify clusters within large datasets.

```
library(knitr)
library(ggplot2)
library(dplyr)
library(plotly)
library(dbscan)
df <- read.csv("pdb_data_no_dups.csv")
```

2.2 Data Summary

The Structural Protein Sequences data ("Structural Protein Sequences," n.d.) frame comprises 14 columns, as indicated by structureId, classification, experimentalTechnique, macromoleculeType, residueCount, resolution, structureMolecularWeight, crystallizationMethod, crystallizationTempK, densityMatthews, densityPercentSol, pdbxDetails, phValue, publicationYear. An examination of the summary statistics revealed the presence of NA and empty values that must be addressed. Additionally, columns deemed irrelevant to our analysis have been removed.

```
head(df, n = 3)
```

```
  structureId   classification experimentalTechnique macromoleculeType
1        100D   DNA-RNA HYBRID     X-RAY DIFFRACTION    DNA/RNA Hybrid
2        101D              DNA     X-RAY DIFFRACTION              DNA
3        101M OXYGEN TRANSPORT     X-RAY DIFFRACTION          Protein
  residueCount resolution structureMolecularWeight
1           20       1.90                  6360.30
```

```
2           24       2.25                  7939.35
3          154       2.07                 18112.80
          crystallizationMethod crystallizationTempK densityMatthews
1 VAPOR DIFFUSION, HANGING DROP                  NA            1.78
2                                                NA            2.00
3                                                NA            3.09
  densityPercentSol                                        pdbxDetails
1             30.89              pH 7.00, VAPOR DIFFUSION, HANGING DROP
2             38.45
3             60.20 3.0 M AMMONIUM SULFATE, 20 MM TRIS, 1MM EDTA, PH 9.0
  phValue publicationYear
1       7            1994
2      NA            1995
3       9            1999
```

```
summary(df)
```

```
 structureId        classification     experimentalTechnique macromoleculeType
 Length:141401      Length:141401      Length:141401         Length:141401
 Class :character   Class :character   Class :character      Class :character
 Mode  :character   Mode  :character   Mode  :character      Mode  :character
```

```
  residueCount        resolution     structureMolecularWeight
 Min.   :     0.0   Min.   : 0.480   Min.   :     314
 1st Qu.:   226.0   1st Qu.: 1.800   1st Qu.:   26129
 Median :   414.0   Median : 2.100   Median :   47478
 Mean   :   825.4   Mean   : 2.264   Mean   :  112079
 3rd Qu.:   820.0   3rd Qu.: 2.500   3rd Qu.:   94085
 Max.   :313236.0   Max.   :70.000   Max.   :97730536
                    NA's   :12812
 crystallizationMethod crystallizationTempK densityMatthews densityPercentSol
 Length:141401         Min.   :  4         Min.   : 0.00   Min.   : 0.00
 Class :character      1st Qu.:290         1st Qu.: 2.21   1st Qu.:44.37
 Mode  :character      Median :293         Median : 2.49   Median :50.50
                       Mean   :291         Mean   : 2.67   Mean   :51.35
                       3rd Qu.:295         3rd Qu.: 2.91   3rd Qu.:57.71
                       Max.   :398         Max.   :99.00   Max.   :92.00
                       NA's   :44362       NA's   :16677   NA's   :16652
 pdbxDetails           phValue       publicationYear
 Length:141401      Min.   :  0.00   Min.   : 201
 Class :character   1st Qu.:  6.00   1st Qu.:2005
 Mode  :character   Median :  7.00   Median :2010
                    Mean   :  6.79   Mean   :2009
                    3rd Qu.:  7.50   3rd Qu.:2014
                    Max.   :724.00   Max.   :2018
```

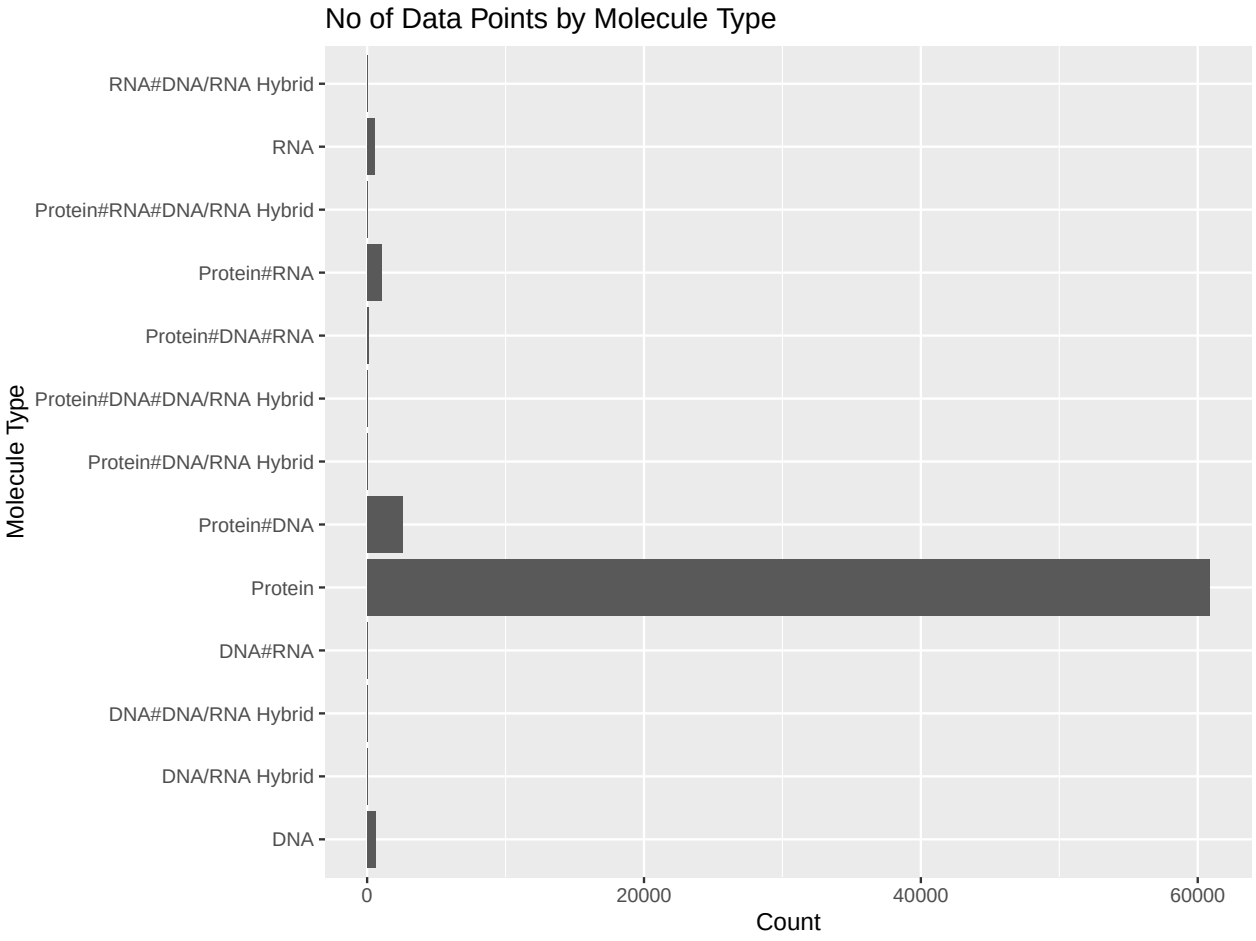
```
                    NA's   :36291    NA's   :23799
```

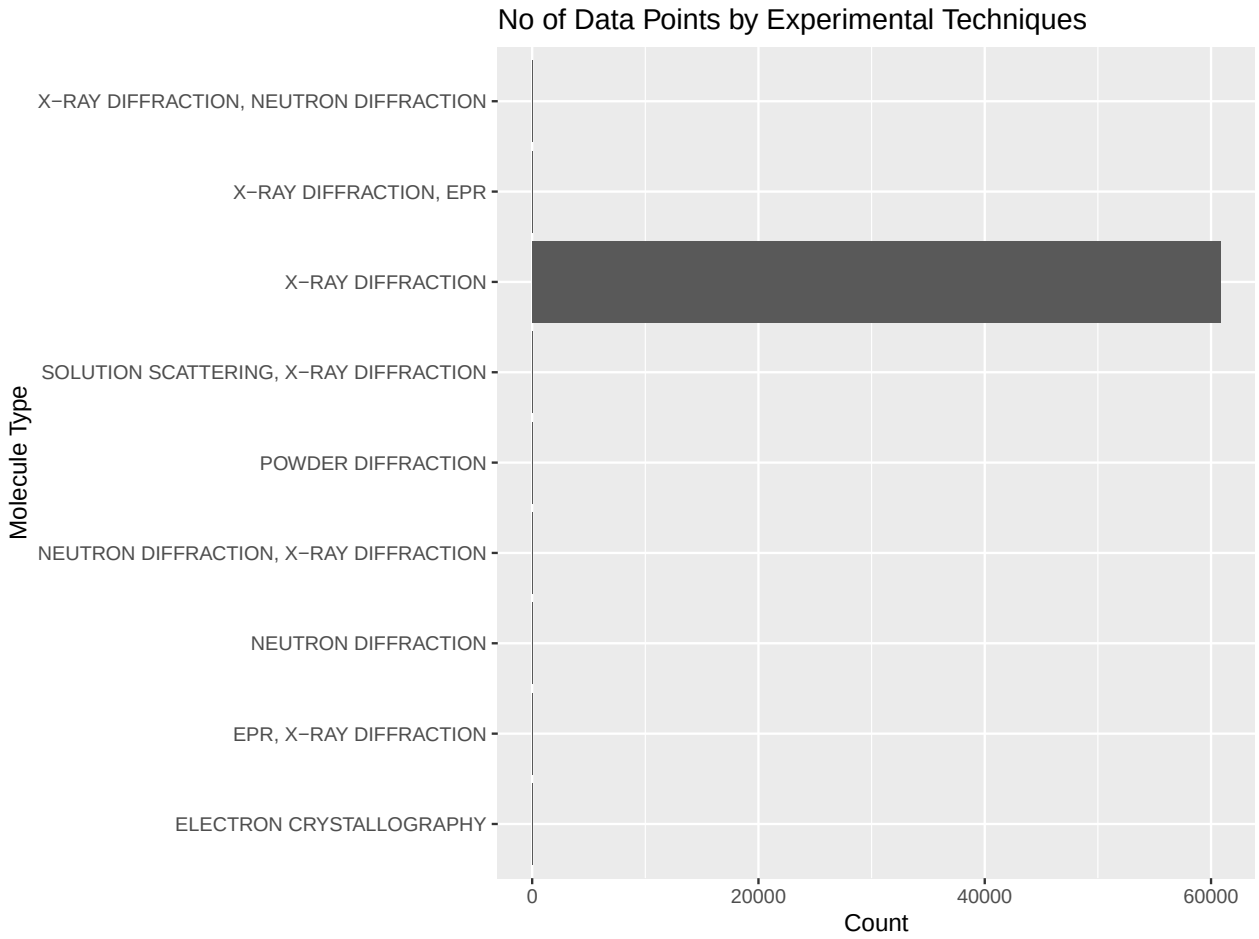
```
df[df == "NULL"] <- NA
df <- df %>% filter_all(all_vars(!is.na(.) & nzchar(.))) %>%
  select(-publicationYear,-pdbxDetails)
```

3 Data Analysis

3.1 Molecule Type and Experiment Technique

First thing that need to be considered are the molecule types and the technique used for the experiment to choose ideal data for examination.





Due to the large amount of data available for proteins, it was selected as the subject of further analysis. Given the majority of the results are obtained through x-ray diffraction, my focus was specifically directed towards proteins that have been studied using this technique.

3.2 Effect of Resolution

Majority of the protein structural discoveries uses the x-ray diffraction and one major achievement in discovery is obtaining a resolution (in angstrom units) as close to one as possible.(Warren 1990; Whittig and Allardice 1986) This is limited by the physics behind it. To understand this one dimensional data I used histogram to see how it is distributed.

The histogram of resolution shows that majority of the experiments can resolve structures with a resolution around 2. (As there are more then 60000 data points in my set, the number of bins is selected as 50 to get the details about the distribution more precisely).

Investigation of structures with resolution below 2 would give much deeper understanding of the structural features and the underlying physics/chemistry.

The number of protein structures with a resolution below two are 26248.

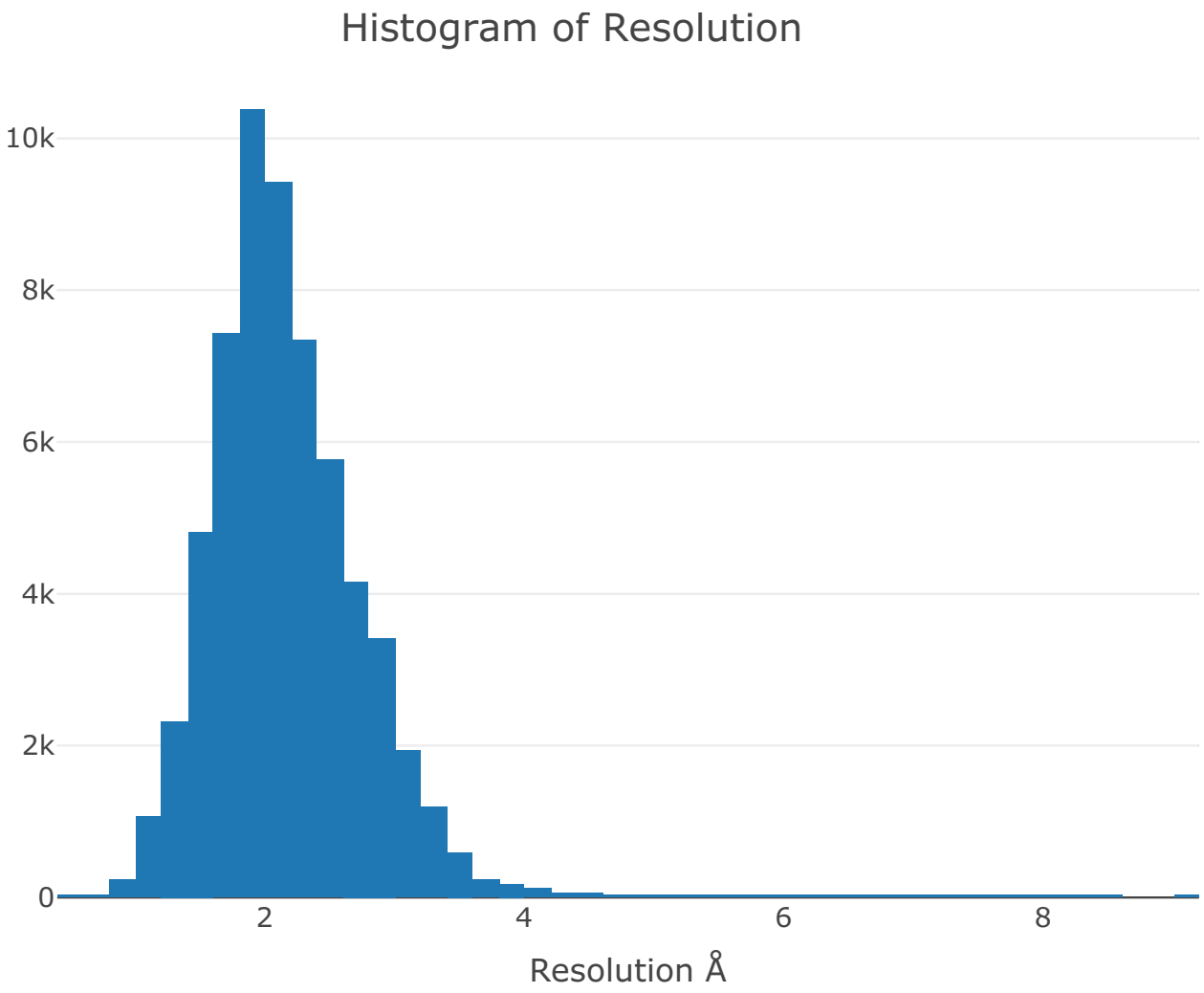


Figure 1: Histogram of Resolution

3.3 Effect of Residue Count

Next I wanted to understand how residue count (the number of residue in macromolecules) affects the resolution of the discovered structure.

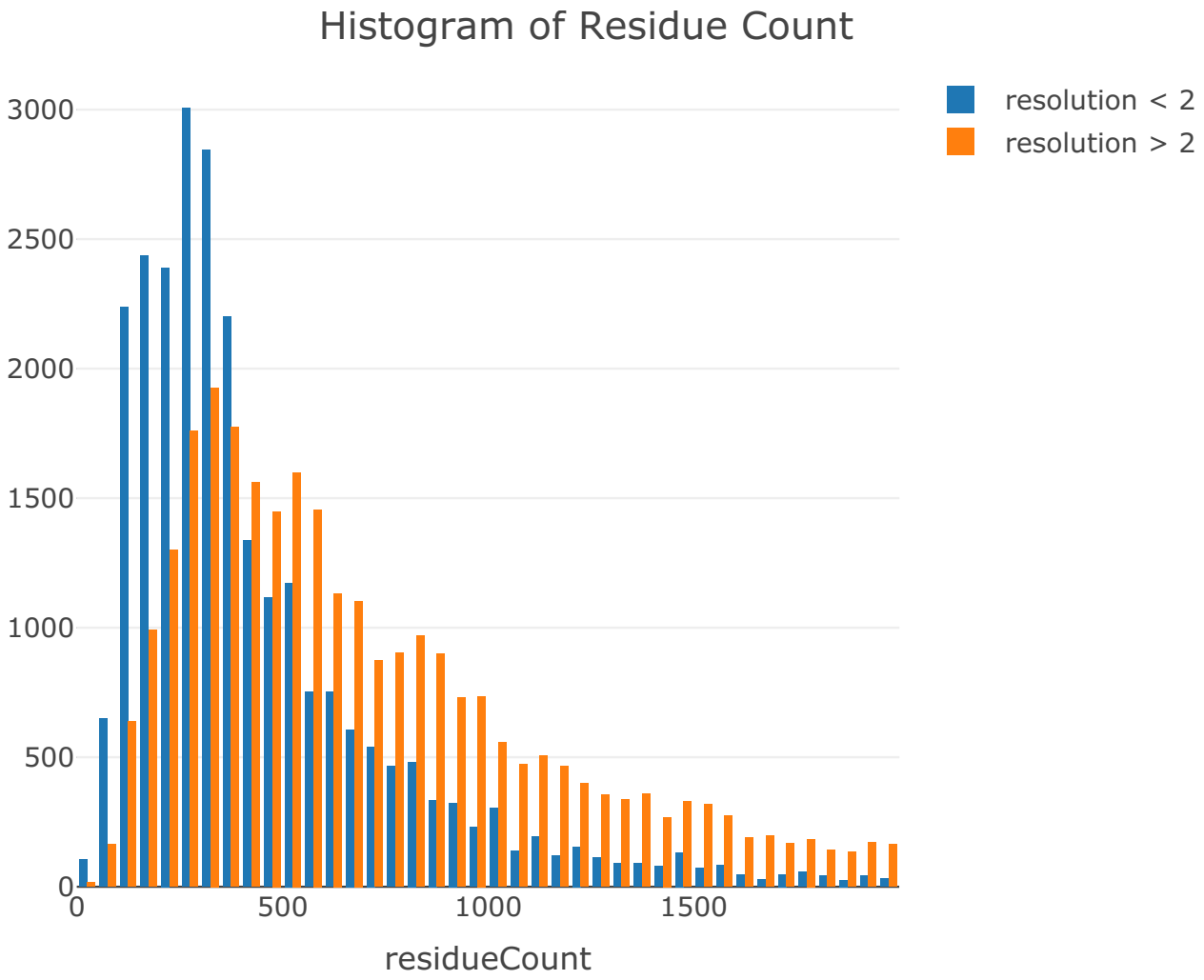


Figure 2: The relationship between residue count and resolution

The relation between resolution and number of residues is intriguing. The plot shows that number of residues in a protein is a factor in determining the protein resolution. There is a negative correlation between the number of residues and resolution. The lesser the number of residues better the resolution (close to 1). Most of the high resolution structures have less than 500 residues. Thus, I investigated that set.

3.4 Exploring Discrepancies in the Density of Proteins in Crystalline and Soluble Forms

Proteins in solution has shown to display features that are being missed in structural discovery. One of the parameters that has been trusted in protein studies in solution is solution density calculation. On the other side structural discovery utilizes a metric depending on the protein crystals. Now the

question is how do proteins act in solution (normal conditions) versus in a crystal lattice (restricted environment) ? Thus I investigated two features from both calculations.

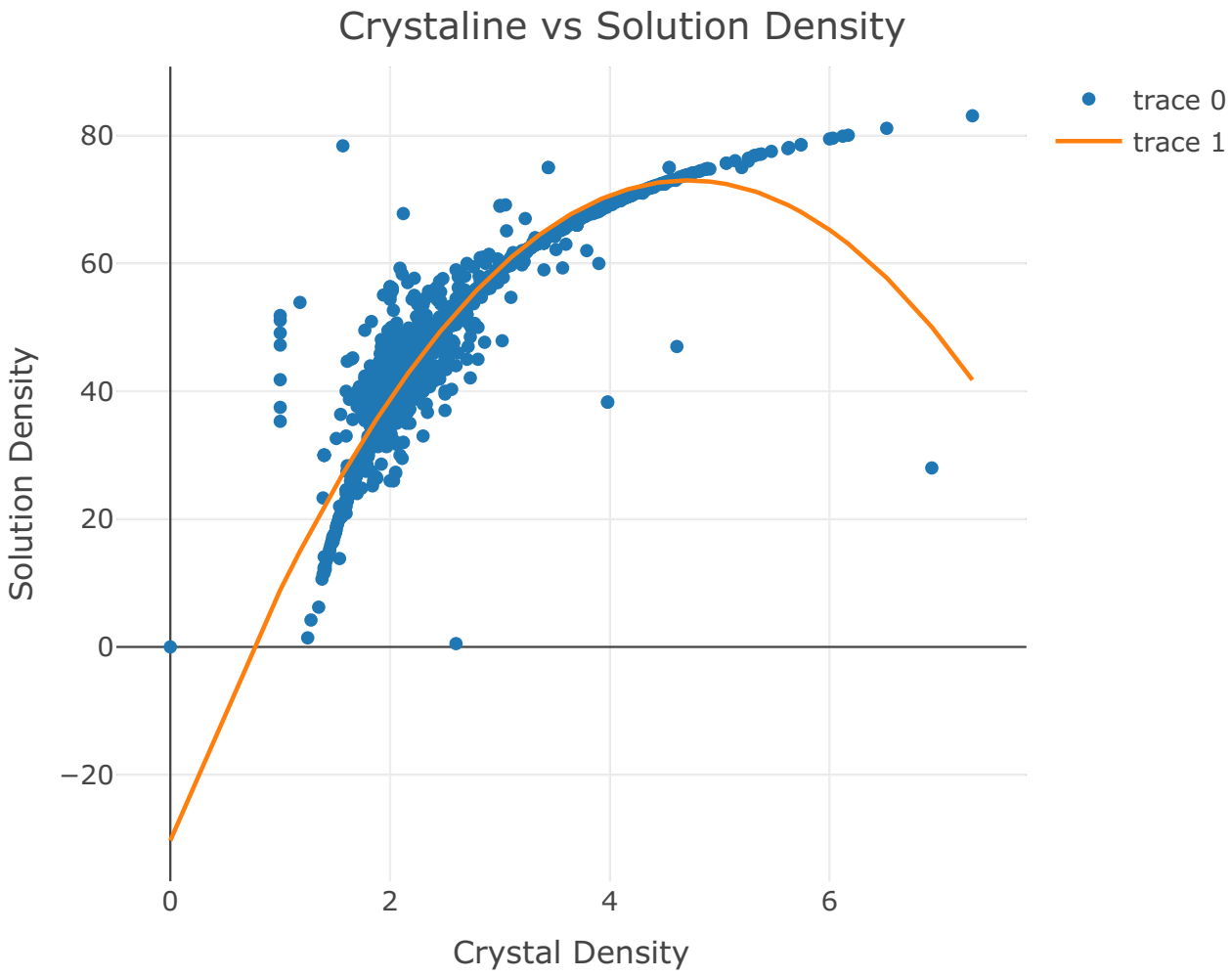


Figure 3: The relationship between crystal density and solution density

The spread of the data in the plot resembles a second degree polynomial fit, hence I added a second degree polynomial fit to capture the underlying principle that correlates these two metrics. However because of the outliers the fit does not recapitulate the data.Thus the outliers needs to be removed.

3.4.1 K-means clustering

```
df_density <- df_residu2 %>% select(densityMatthews,densityPercentSol)
kmeans_fit <- kmeans(df_density,10,iter.max = 20, nstart = 2)
df_density$cluster <- as.character(kmeans_fit$cluster)
```

I used K-means clustering to discover similar data points in the clusters within the two feature space. It can reveal hidden patterns and trends such that it can cluster and separate the data points that don't follow the correct trend.There are thousands of data points in my system therefore I selected k value as 20 to be able to generate unique clusters.

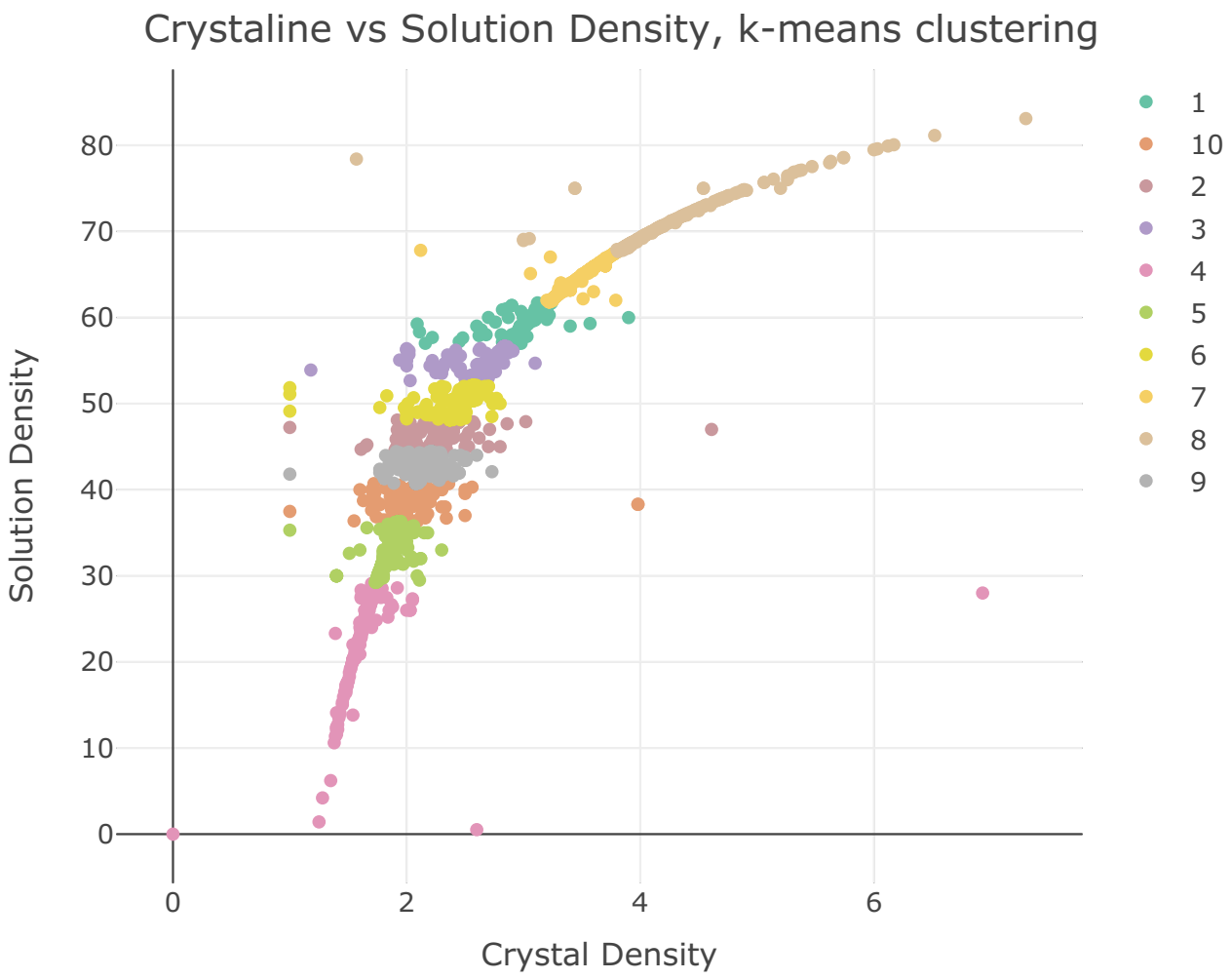


Figure 4: K-means clusturing to analyse relationship between crystal density and solution density

The K-means clustering failed to identify the correct clusters, therefore I used Hierarchical clustering that can identify clusters using agglomerative method which does not require a prior specification of the number of clusters.

3.4.2 Hierarchical clustering

```
df_density2 <- df_residu2 %>% select(densityMatthews,densityPercentSol)
dist_mat <- dist(df_density2, method = "euclidian")
hclust_fit <- hclust(dist_mat, method = "ward.D2")
df_density2$cluster <- as.character(cutree(hclust_fit, k = 20))
```

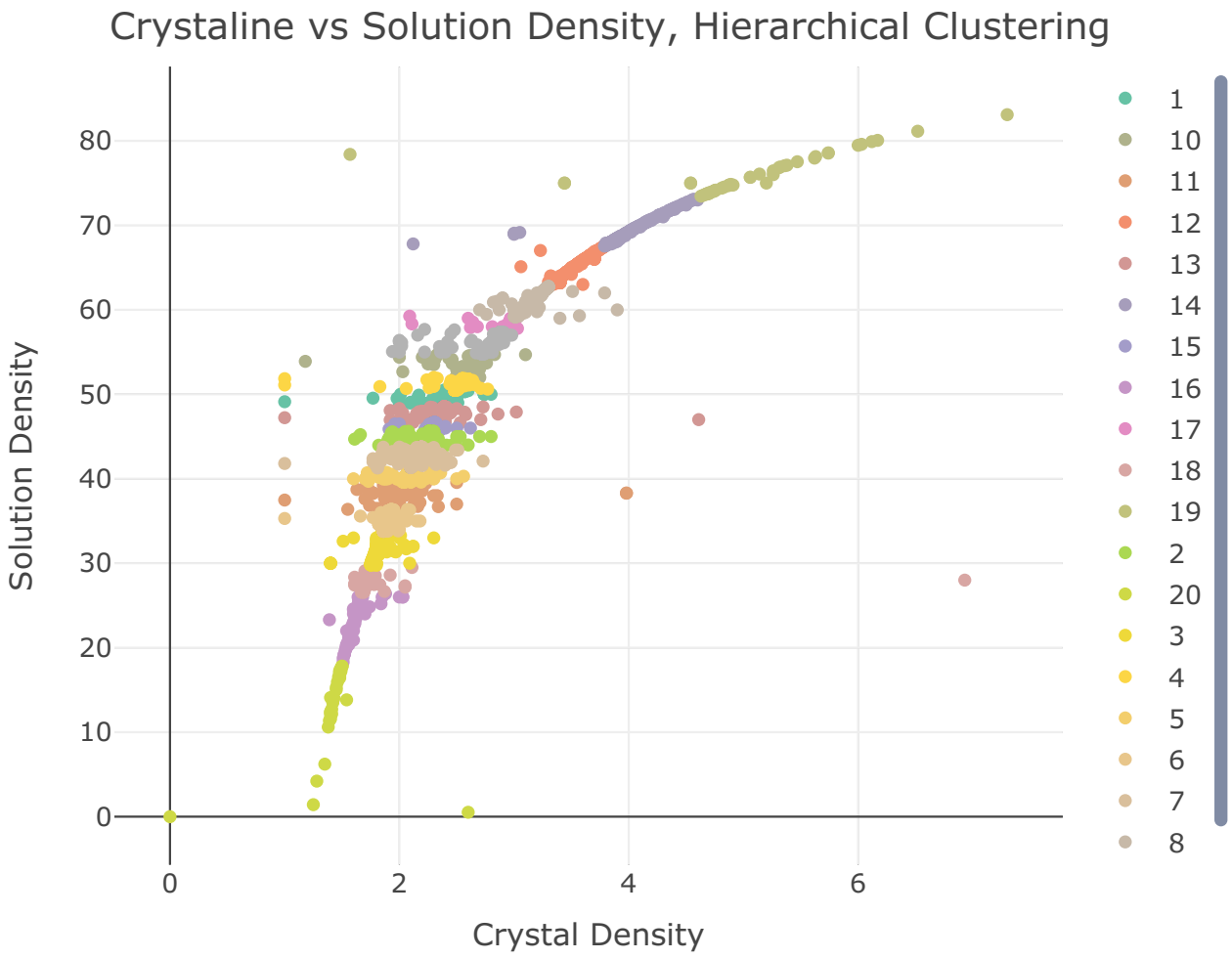


Figure 5: Hierarchical clustering to analyse the relationship between crystal density and solution density

The figures indicate that kmeans and hierarchical clustering was not the best approach to identify the main overlapping points versus the outliers in the data. This data show dense regions around several data points. Therefore a density based clustering approach would be more useful in understanding this data. To achieve that I used dbscan function. DBscan is a density-based clustering

algorithm that groups together data points that are close in space and separates points that are far apart. Dbscan is useful because it can find clusters of arbitrary shapes and sizes, unlike k-means and hierarchical clustering which tend to find circular clusters.

3.4.3 Density Based Clustering

```
df_density3 <- df_residu2 %>% select(densityMatthews,densityPercentSol)

# Perform DBSCAN on the df_density3 data frame
result <- dbscan(df_density3, eps = 0.1, MinPts = 1)
df_density3$cluster <- result$cluster
```

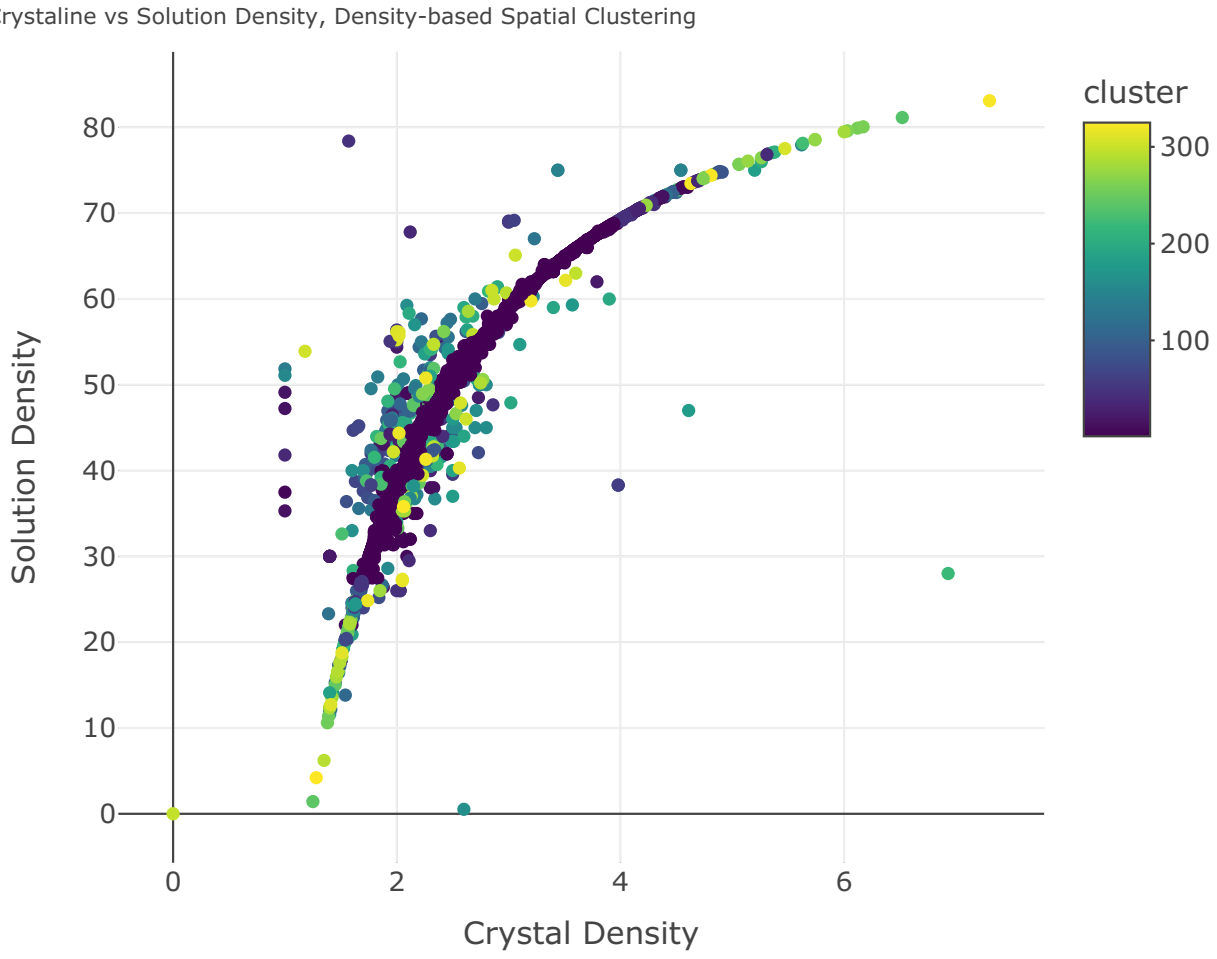
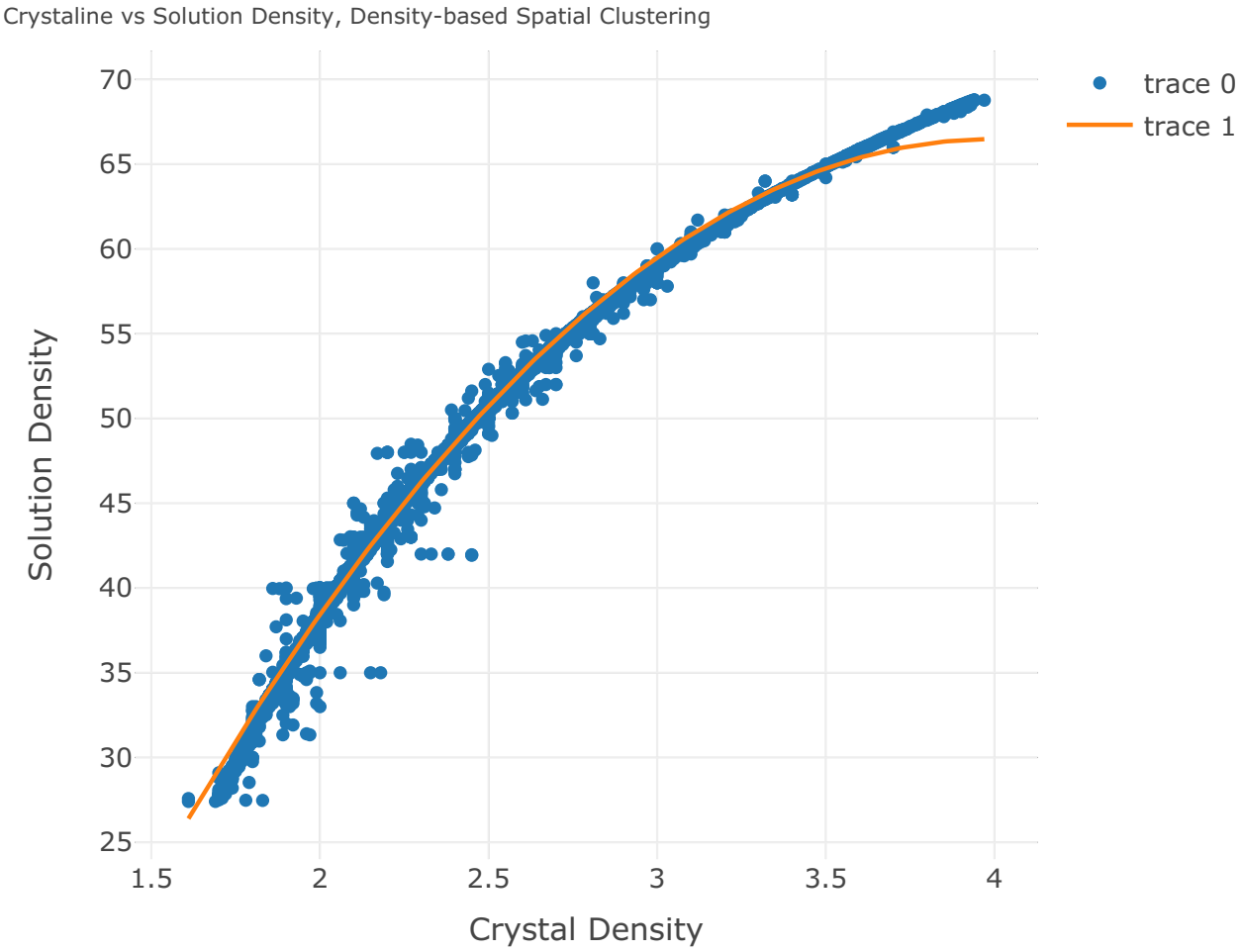


Figure 6: Density-based Spatial Clustering to analyse relationship between Crystaline vs Solution Density

Looking at the plot we can clearly see the outliers in our data set. To focus more clearly on the data I choose main cluster and re-plot the data. I added a second degree polynomial fit once again to discover the underlying relation. The polynomial fit now captures more than 95% of the data as shown in the plot below.



4 Conclusion

The results of this analysis demonstrate that through the utilization of various machine learning techniques, it is possible to identify and decipher fundamental relationships within a data set. This study highlights the challenges of achieving results that resemble natural phenomena using these techniques, with only 26% of the data points, equivalent to only 26% of the experiments, resulting in naturally occurring behavior that captures the correct underlying fundamentals.

5 References

"Structural Protein Sequences." n.d. https://www.kaggle.com/datasets/shahir/protein-data-set.

Warren, Bertram Eugene. 1990. *X-Ray Diffraction*. Courier Corporation.

Whittig, L. D., and W. R. Allardice. 1986. "X-Ray Diffraction Techniques." *Methods of Soil Analysis: Part 1 Physical and Mineralogical Methods* 5: 331362.