**Data Analysis of Binding Affinity of Biomolecules**

**Data Retrieval**

Data for this project was found on the RCSB Protein Data Bank site found at <https://www.rcsb.org/>.

Data for DNA and RNA sequences were filtered on the site and then downloaded to two csv files. Data for Proteins were also filtered on the site and then downloaded into three csv files.

**Part One (File merging)**

The DNA and RNA csv files were read into a Jupiter notebook and concatenated to create one large file. Non-pertinent columns were removed.

Part One (Analysis) -

The file created in the Part One file merge was read into a new Jupiter notebook. The file was cleaned by removing any rows containing all NaN values. The Release Date column was changed to DateTime format.

We created two new data frames, one containing only rows for the Experimental Method “Solution NMR” and the other for the Experimental Method “X-Ray Diffraction.” From the original data frame we found the range of Release Date and then created bins to group the data frame by years. We then created a bar chart representing the number of sequences found in each year’s grouping. We created a data frame that showed the number of sequences found in each year for the two Experimental Methods and then plotted that bar chart. We then grouped the original data frame by structure type of DNA, RNA or Hybrid and created a bar chart showing the number of sequences by year group.

Finally, we created three box plots and found any outliers for each of the three structure types comparing their size by Molecular Weight.

**Part Two (Analysis)**

To continue the analysis, the same merged data file was used to clean the data for RNA and DNA values.

With the clean data, we found the Ligands corresponding to DNA, RNA only. These data frames were then used to find the Ligands with minimum binding affinity (Kd). As lower the value, the better the affinity. Box plots helped us to find the outliers in data for binding affinity. The outliers were removed to help us further investigate the Linear correlation between Molecular weight and Value. The scatter plot helped to find the correlation between Ligand and value. To conclude we calculated the r2 value for DNA and RNA. The conclusion was made that there is no significant correlation between the molecular weight and binding affinity.

**Part Three (Analysis)**

In this part of the analysis, data for protein and its ligands was obtained from the PDB database as 3 separate CSV files. These files were loaded into pandas, and a data frame was created for only specific columns and using the concat function, merged into one data frame. Outliers were found and omitted and a regression was performed between the size (MW) of the proteins and their binding affinities. CSV files for DNA and RNA ligands were loaded into a data frame from part two of the analysis, and an outer merge was performed on the three data frames to find the common ligands. A chart with binding affinities of the same ligand for different biomolecules was plotted to show the results.