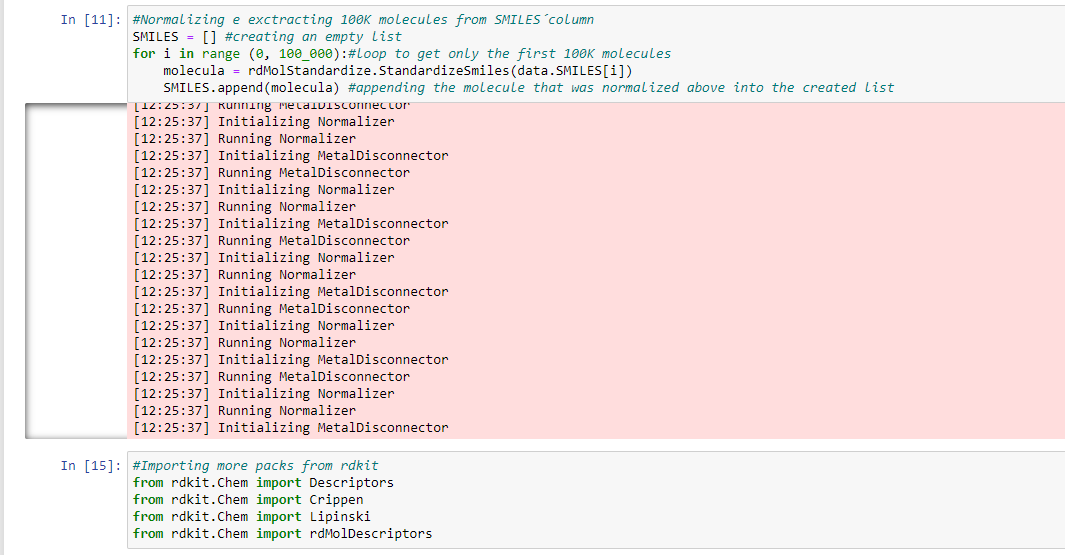
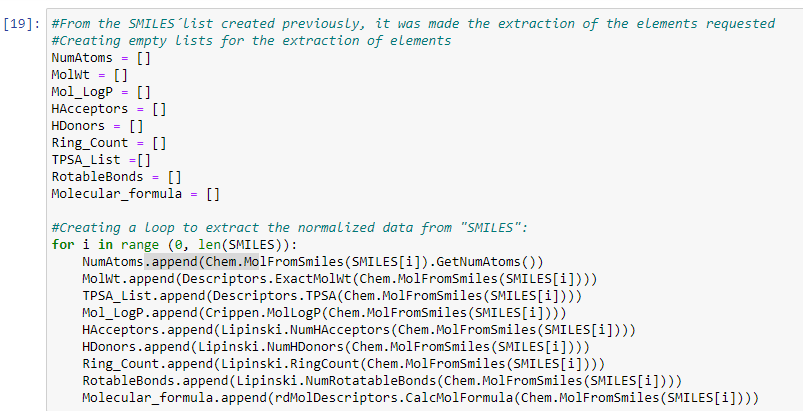
ZINC 12

This project was proposed by our R Introduction for Data Science´s teacher [Wandré Nunes de Pinho Veloso](https://www.linkedin.com/in/wandreveloso/?originalSubdomain=br) and it was made by two students in the second semester of Data Science - CEUB: [Henrique de Andrade Araujo](https://www.linkedin.com/in/henrique-de-andrade-araujo-56467123a/) and [Matheus Lopes Amado](https://www.linkedin.com/in/matheuslopds/). The task was to work with a chemical compound library so we could learn how to manipulate data and create important graphs on it. Especifically, we had to select the first 100,000 compounds and work with their SMILES, Molecular Formula, ExactMolecular Weight, Molecular LogP, Number of Hydrogen Acceptors and Donors, Number of Rotatable Bonds, Ring Count and TPSA. Python and R were used on this project.

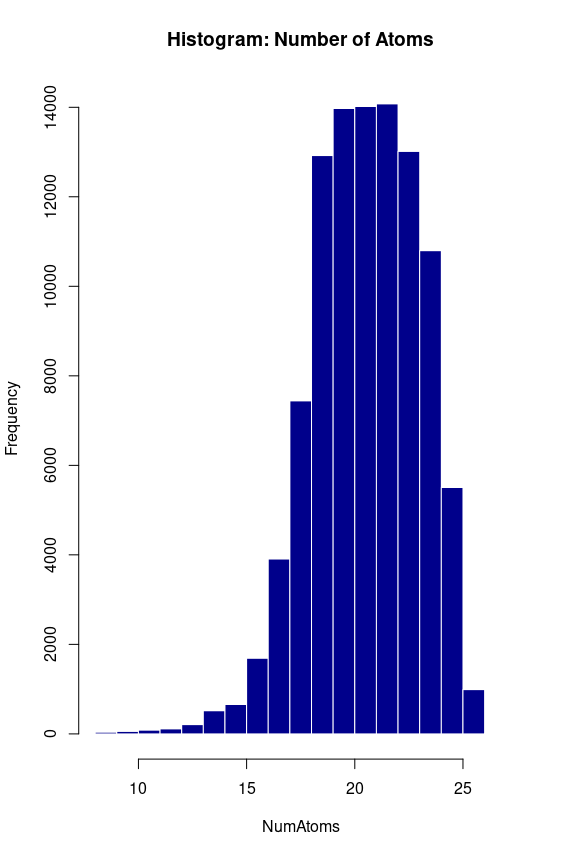
The database that was chosen for us was ZINC12. ZINC is a free database of commercially-available compounds for virtual screening. The version that was used on this project (ZINC12) was created in 2011 and released to the public in 2012. It stopped being supported in 2015. Currently, there is a better and more updated version called ZINC15 that has been supported since 2015. The ZINC database contains about 35 million 3D molecules compounds provided in ready-to-dock. ZINC is free and provided by the Irwin and Shoichet Laboratory in the Department of Pharmaceutical Chemistry at the University of California San Francisco (UCSF). This library was first designed with the purpose of targeting based virtual screening, and according to current holders/owners, it is still the case. It is also used in other applications like finding compounds to purchase, finding compounds annotated and predicted for a particular target, among others.

As it was said previously, R and Python were the programming languages we used on this project. First, we downloaded the archive as .csv and imported it to python by using pandas. Later on the project, we also used “rdkit”. We normalized the data of 100,000 molecules using “rdkit” resources and we got rid of NAs.

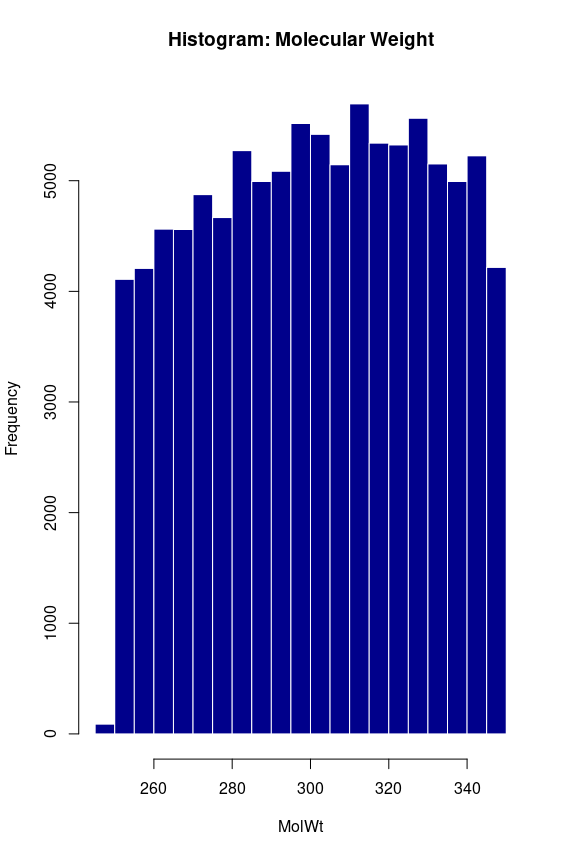


Continuing, we created a dict to replace the labels on the csv. After that, we started using R (Posit Cloud) to get graphics and conclusions based on that dataframe.

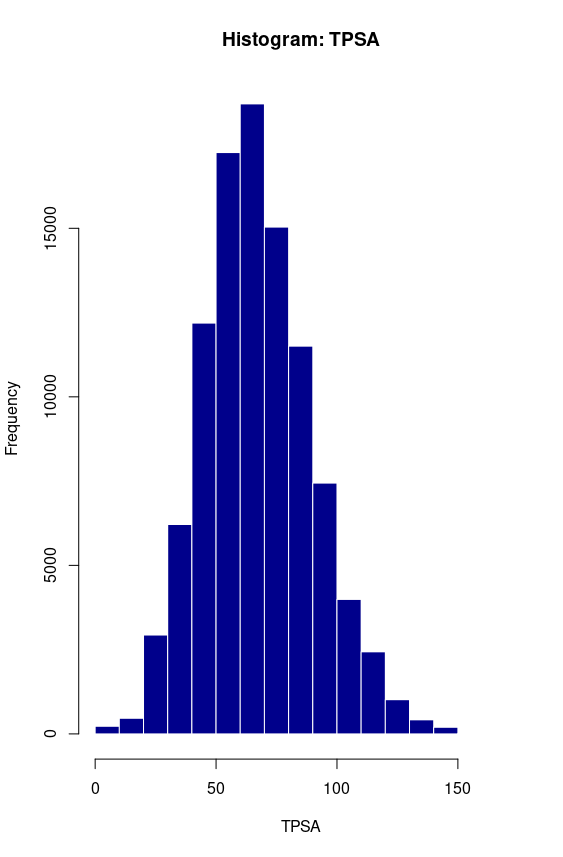
In R, we used three libraries: “tidyverse”, “ggplot2” and “dplyr”. First we uploaded the csv archive to R and read it. Then, we decided to explore the **Number of Atoms** on these molecules. We came to the conclusion that it varies between 8 and 26 atoms per molecule. The mean is = 21.01 atoms per molecule and the median = 21. From the histogram we noticed that most of the molecules have between 18 and 21 atoms.



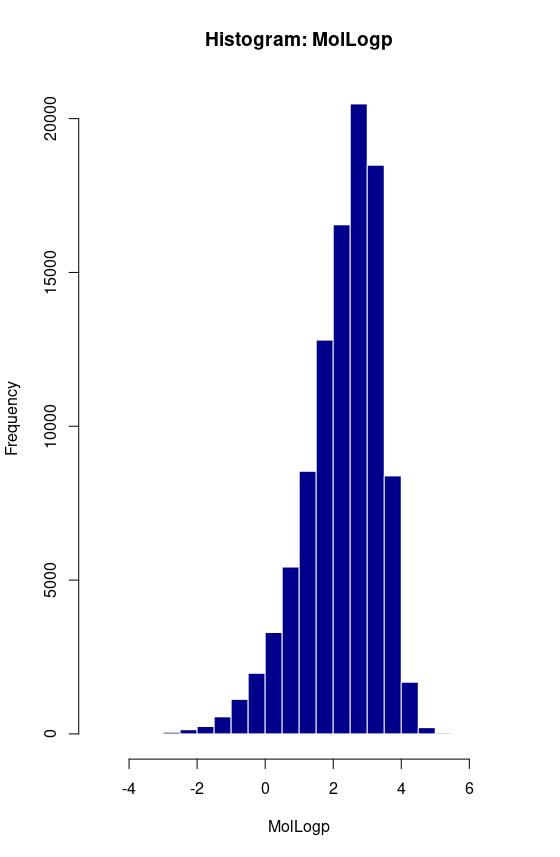
Second, came **Molecular Weight.** From what we abstracted, it varies between 250 and 350 grams/mole. The mean is = 300.88g/mol and the median = 301.14g/mol. Most of the molecules weigh between 285 and 330 grams/mol. As it can be observed in the graph below:



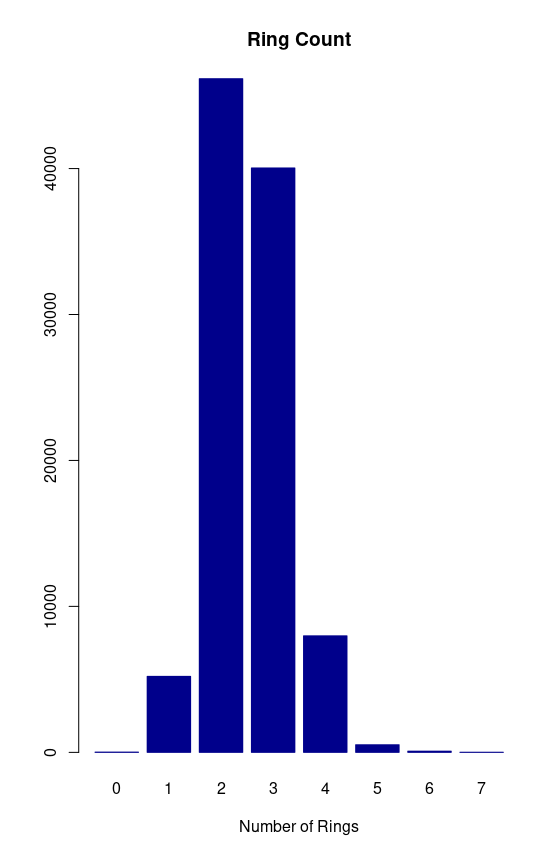
Next, we explored the **TPSA (Measure of the Polar Surface Area)**, and it varies from 0 to 155 units of measure(?). The mean is = 67.45 and the median = 66.04. The graph below shows that most molecules are concentrated where the TPSA is 50 and 70 units of measure(?).

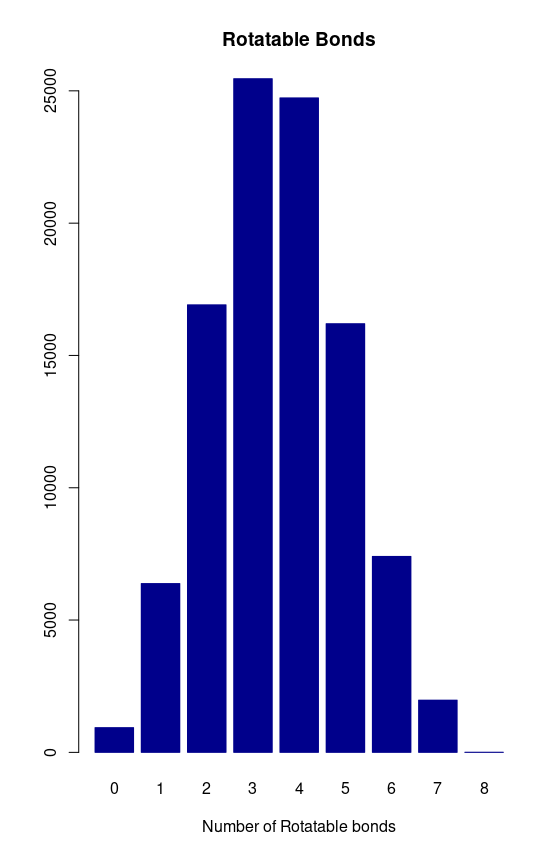


Following, we analyzed the **Logp of the Molecules (MolLogp)** and it varies from -5 to 6 units of measure(?). The mean is = 2.28 and the median = 2.48. The MolLogp measures how hydrophobic a molecule is**.** Majority is between 1.5 and 3.5 units of measure

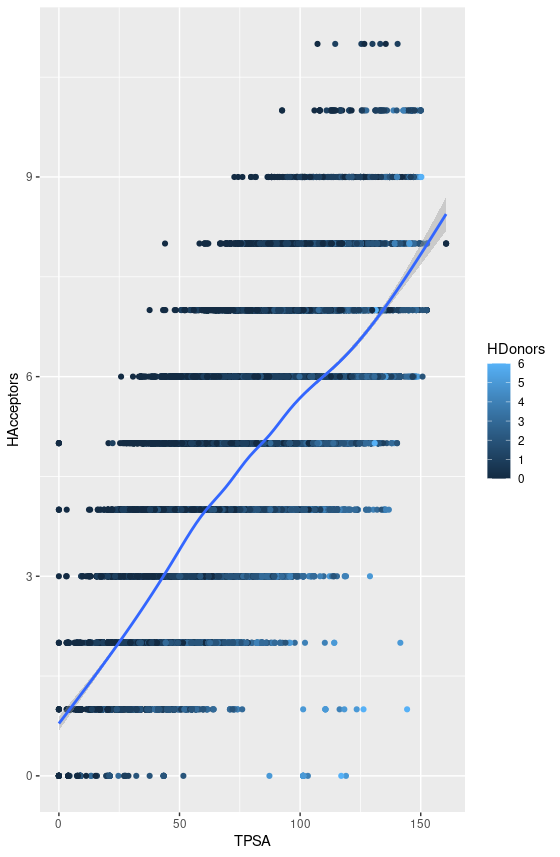


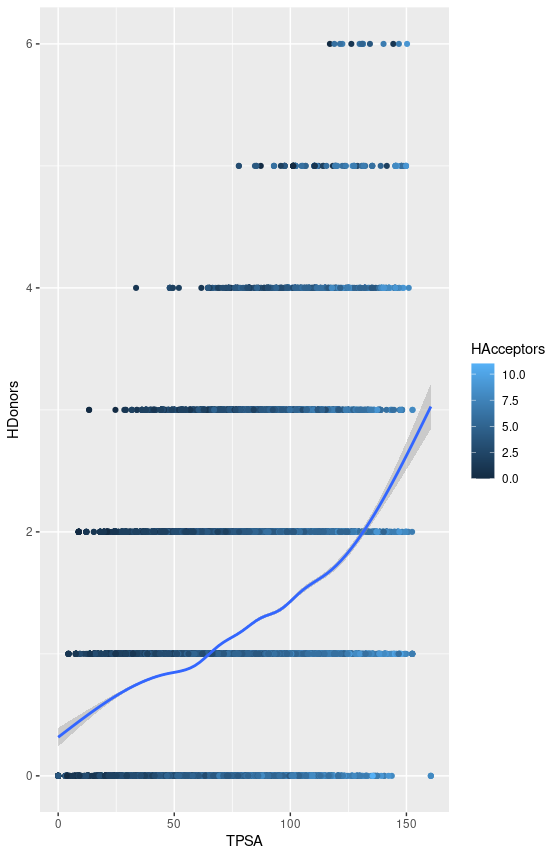
Once the continuous quantitative variable was explored, we looked at the discrete part of it and observed that there are: **amount of Rings**, regions who donate and accept H, and the **Rotatable bonds** per molecule. The number of rings in each individual molecule varies from 1 to 7, and most of them (86%) have 2 or 3 rings. The number of rotatable bonds varies from 0 to 8. The majority of the molecules (25%) have 3 rotatable bonds, followed by 4 rotatable bonds (24%), and the molecules with 2 and 5 rotatable bonds represent 16% each.





Last but not least, the H acceptor and donor regions vary from 0 to 11 and from 0 to 6 respectively. The number of H acceptors regions in a molecule has strong correlation (0.7) with TPSA, on the other hand, the number of H donors regions has a weak correlation (0.3) with TPSA.





Fontes:

<https://zinc12.docking.org/>

<https://wiki.bkslab.org/index.php/ZINC12>