**CENTRO DE ENSINO UNIFICADO DE BRASÍLIA – CEUB**

**CURSO DE BACHARELADO EM CIÊNCIAS DE DADOS E *MACHINE LEARNING***

**Introdução a R para Ciência de Dados**

**Professor Dr. Wandré Nunes de Pinho Veloso**

**Alunos: Gabriel Assis de Paula, RA: 22201568**

**Yasmin Taube dos Santos, RA: 22204365**

**TRABALHO PRÁTICO:**

**Caracterização de biblioteca de compostos químicos.**

**Brasília, Novembro de 2022**

O trabalho tem como objetivo caracterizar uma biblioteca de compostos químicos utilizando a linguagem de programação R. A biblioteca escolhida para a análise foi a disponibilizada pelo DrugBank através da sua plataforma online oficial, <<https://go.drugbank.com>>.

O DrugBank (mantido e atualizado pela Universidade de Alberta, em Edmonton, Canadá) é um banco de dados iniciado em 2006 que apresenta informações detalhadas sobre milhares de compostos químicos e farmacológicos de medicamentos, sendo referência para profissionais da área da química, saúde, estudantes e para o público em geral. A versão 5.1.9 empregada nesta análise, dispõe de 15.085 compostos e é também a mais recente, até o momento, lançada no dia 03 de Janeiro de 2022.

**Análise de dados**

Após o download da base de dados, foi utilizado o pacote de ferramentas para quimioinformáticaRDKit através da linguagem Python para gerar um arquivo CSV contendo os cálculos dos seguintes descritores moleculares:

* **SMILES**

(Simplified Molecular Input Line Entry System) é uma notação química voltada à computação.

* **Molecular\_formula**

Fórmula molecular do composto.

* **ExactMolWt**

Massa molar exata dos compostos calculada em g/mol (gramas por mol).

* **MolLogP**

Coeficiente de partição.

* **RingCount**

Quantidade de anéis do composto.

* **NumHAcceptors**

Quantidade de átomos de hidrogênio que o composto pode receber.

* **NumHDonors**

Quantidade de átomos de hidrogênio que o composto pode doar.

* **NumRotatableBonds**

Quantidade de anéis rotativos do composto.

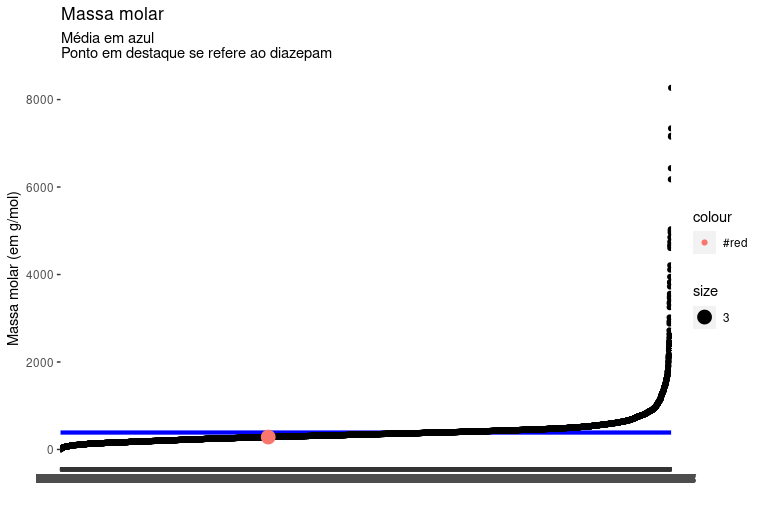
* **TPSA** (Total Polar Surface Area)

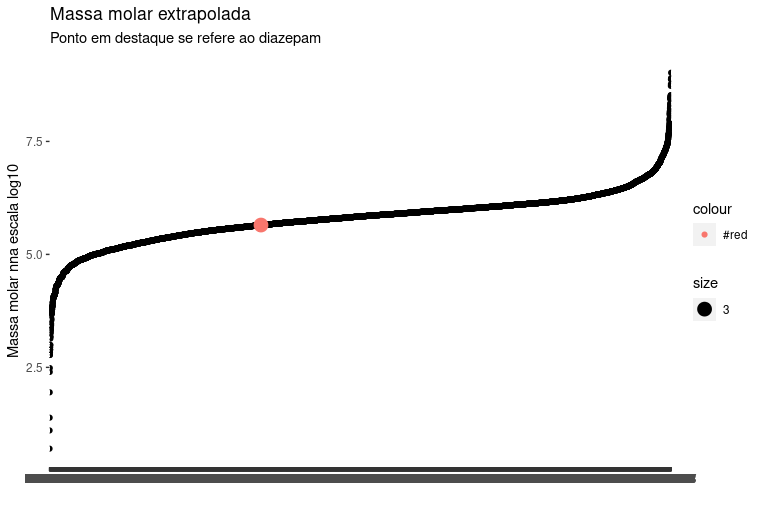
soma da superfície dos átomos polares do composto

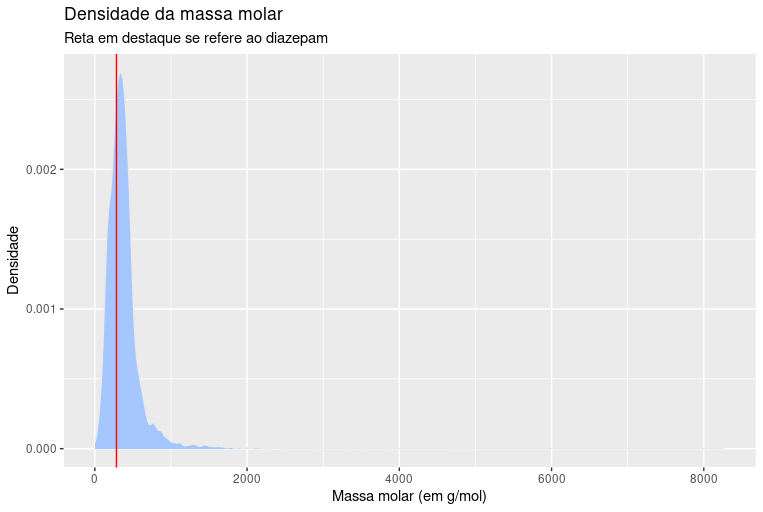
Lendo o arquivo CSV através da linguagem R foi possível inferir as seguintes análises utilizando os pacotes *dplyr*, *ggplot2* e *tidyverse*:

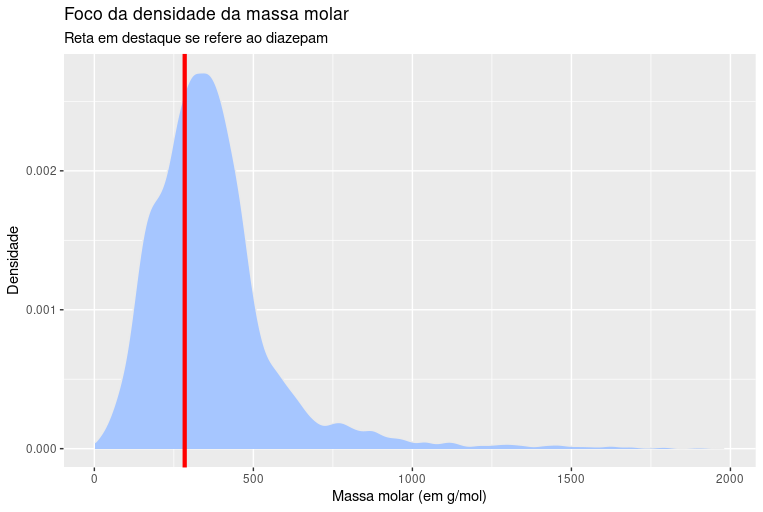
* **ExactMolWt**

Média de massa molar dos compostos: 387,624 g/mol; Peso do composto mais leve é de 2,016 g/mol; Peso do composto mais pesado é de 8268,351 g/mol.



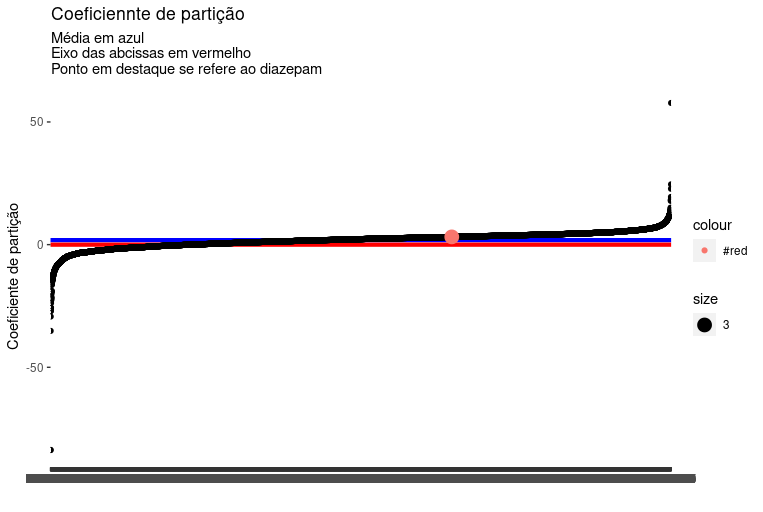




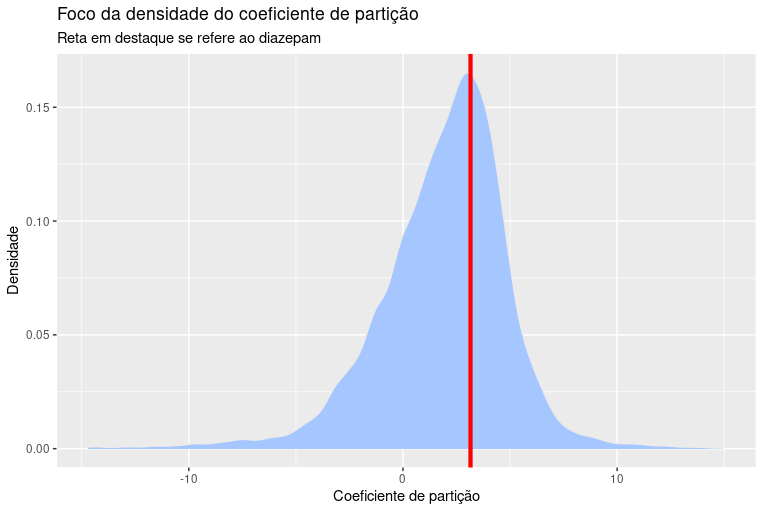


* **MolLogP**

Média do coeficiente de partição: 1,8278; O menor coeficiente de partição é de -83,7536; O maior coeficiente de partição é de 57,75100.

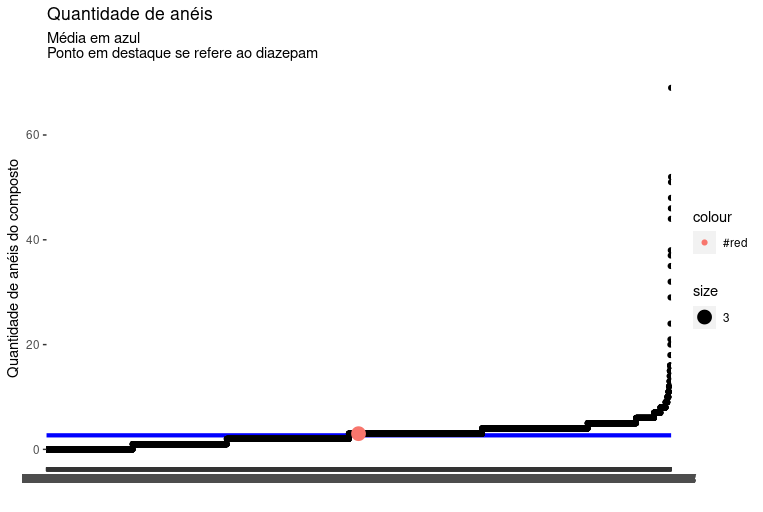


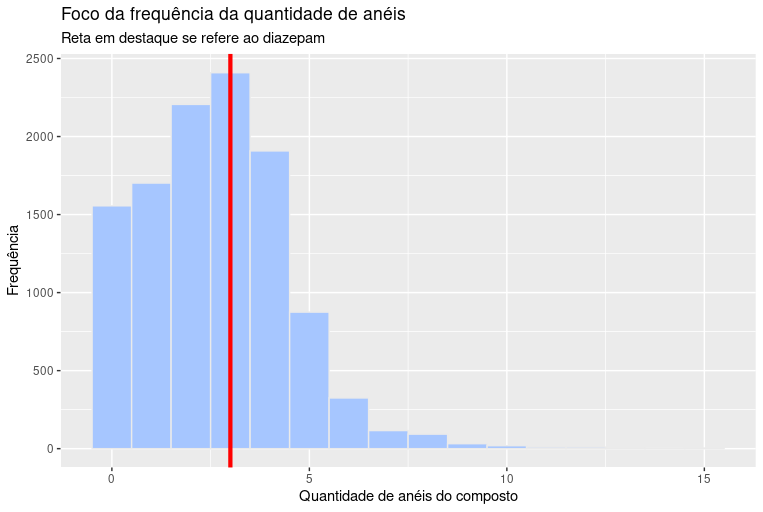
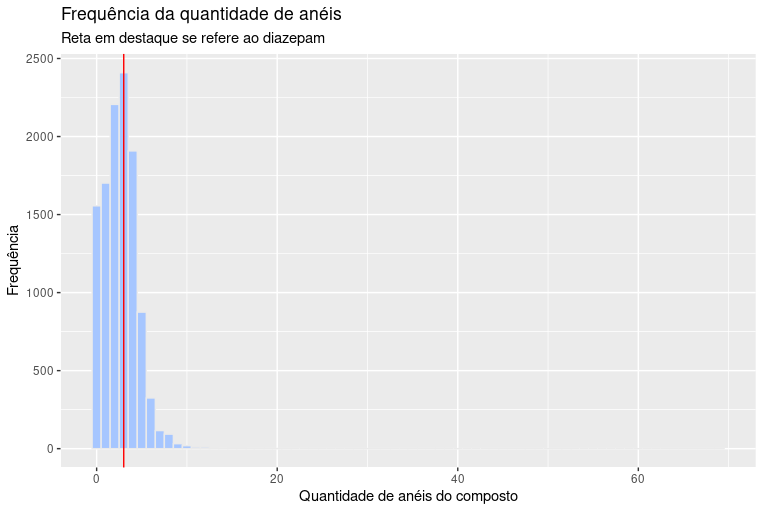




* **RingCount**

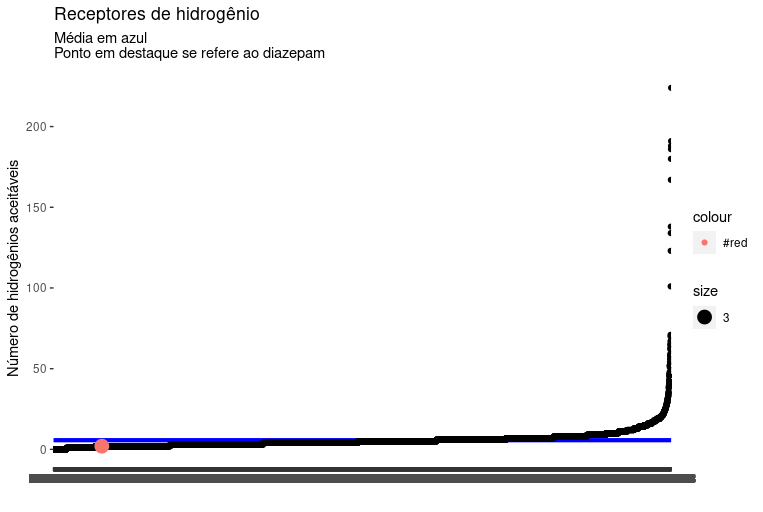
Média de anéis: 2,687; Menor quantidade: 0; Maior quantidade: 69.

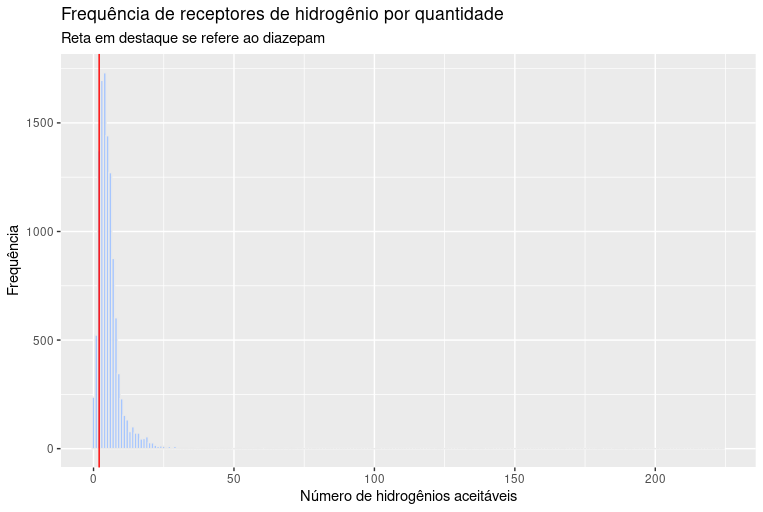


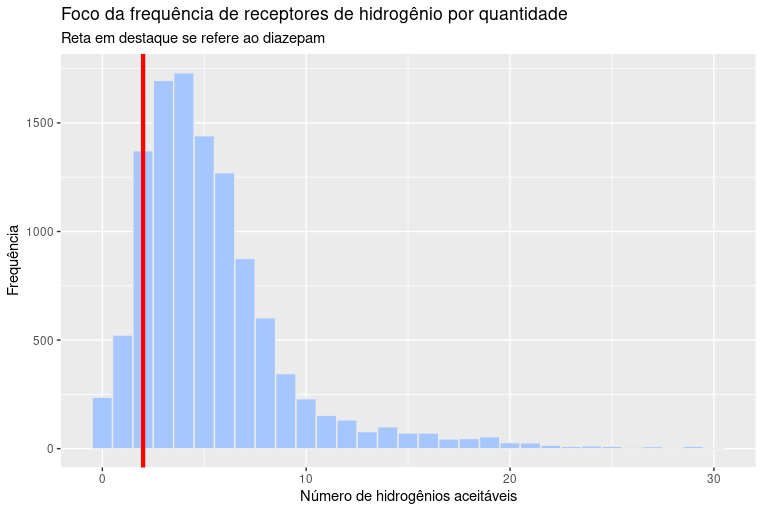


* **NumHAcceptors**

Média de aceptores: 5,709; Menor quantidade: 0; Maior quantidade: 224.

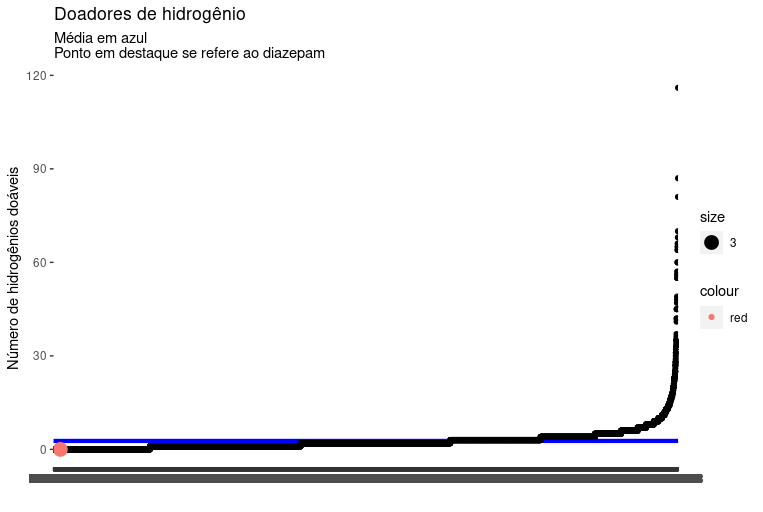


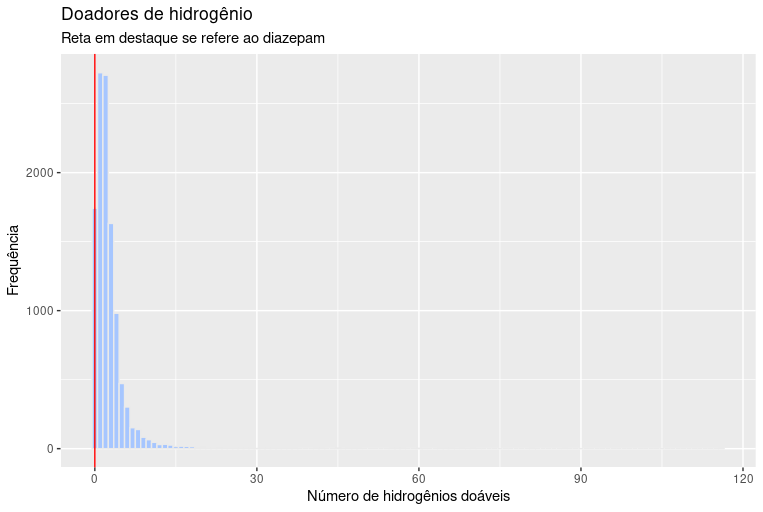


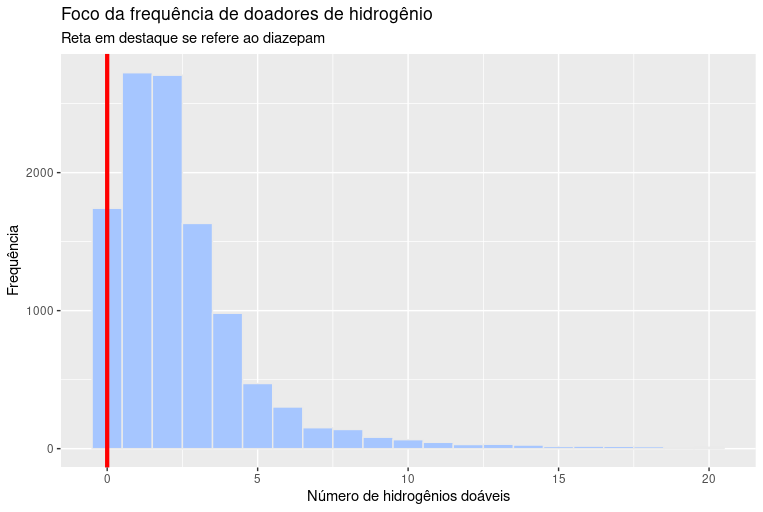


* **NumHDonors**

Média de doadores: 2,722; Menor quantidade: 0; Maior quantidade: 116.

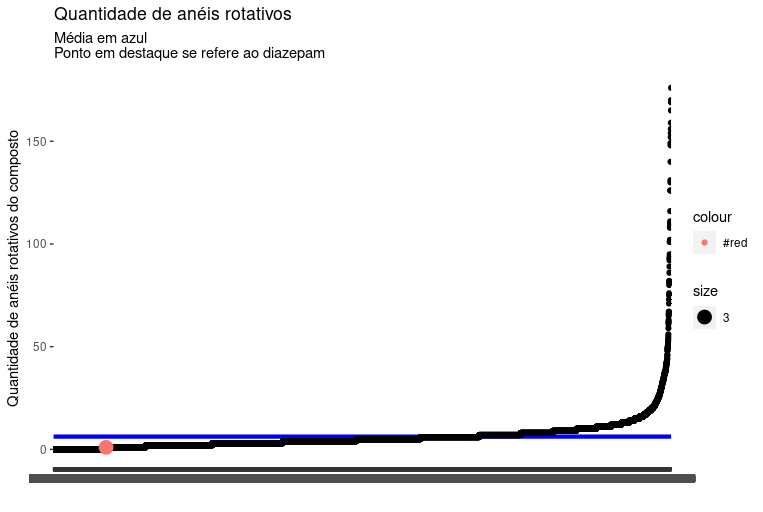




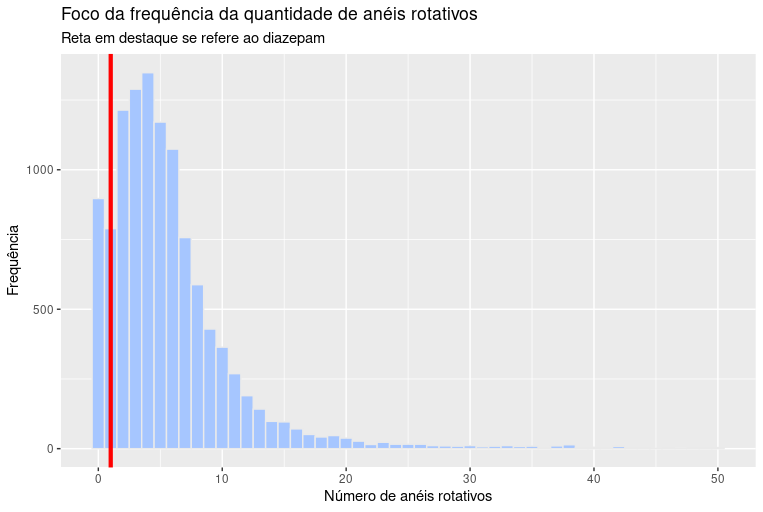


* **NumRotatableBonds**

Média de anéis rotativos: 6,236; Menor quantidade: 0; Maior quantidade: 176.

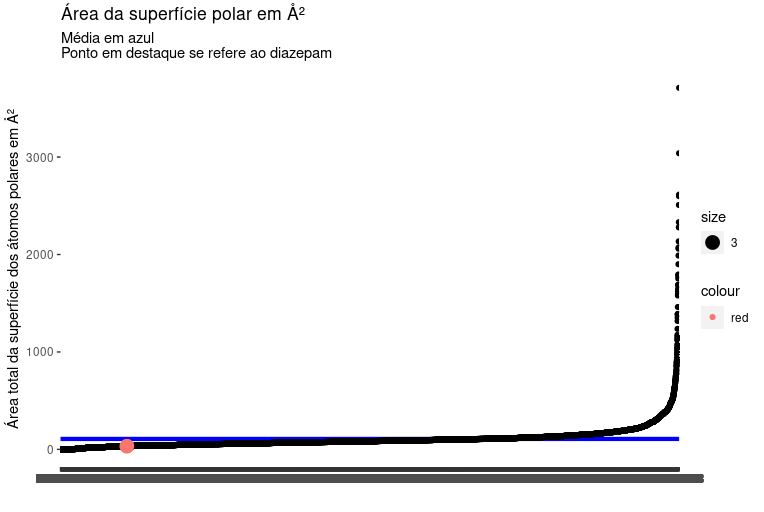


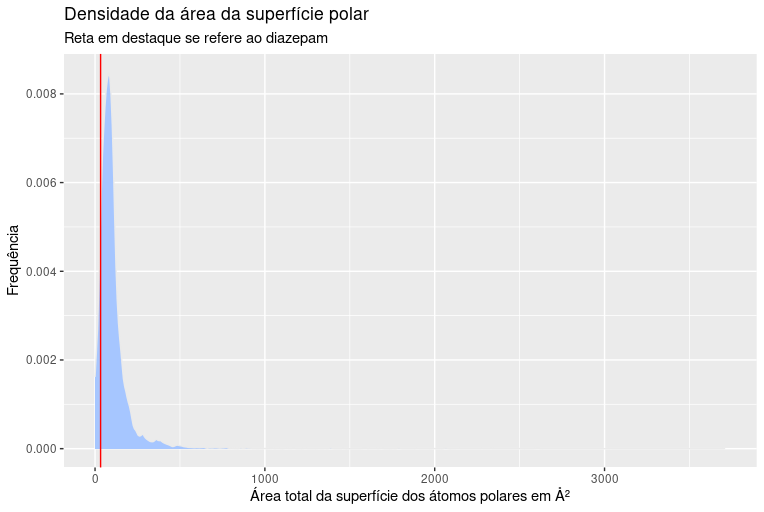


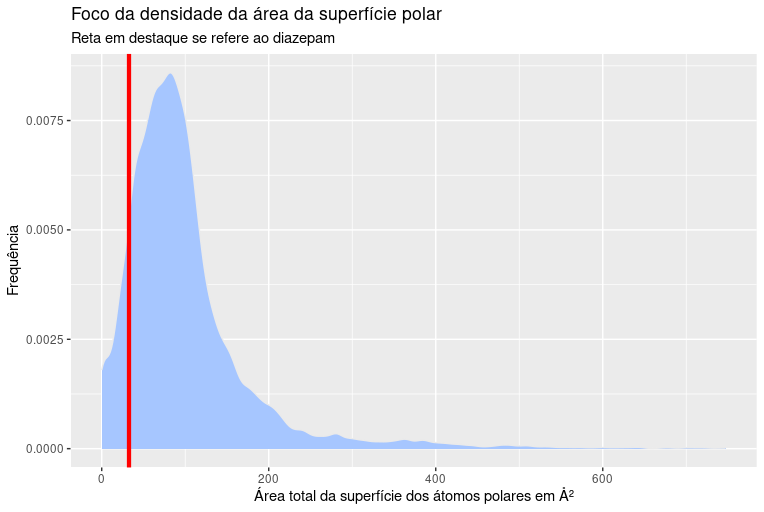


* **TPSA**

Média das áreas polares: 107,26; Menor quantidade: 0; Maior quantidade: 3710,38.







Outras colunas presentes no CSV são:

* **Database.ID**

ID do composto na base DrugBank.

* **Drug.Groups**

Status atual do composto em relação ao seu uso medicinal.

São esses:

* *approved*

oficialmente aprovados para serem comercializados (pode variar com a legislação do país)

* *vet\_approved*

composto aprovado para uso em animais.

* *illicit*

drogas ilícitas (pode variar com a legislação do país).

* *experimental*

droga experimental em fase de descobrimento ou pré- descobrimento.

* *investigational*

droga em fase de desenvolvimento e triagem científica.

* *withdrawn*

costumava ser comercializada mas foi descontinuada ou proibida (pode variar com a legislação do país).

* *neutraceutical*

composto regulado a nível farmacêutico e possui efeito nutricional.

* **Generic.Name**

Nome científico ou genérico do composto.

* **Products**

Lista de produtos fármacos que utilizam o composto (o DrugBank usa apenas como referência os mercados dos Estados Unidos, Canadá e Europa).

**REFERÊNCIAS BIBLIOGRÁFICAS**

Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2017 Nov 8. doi: 10.1093/nar/gkx1037.