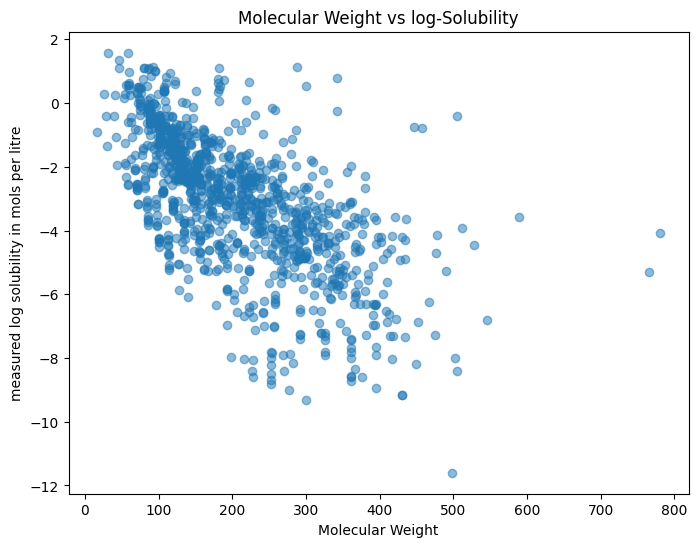
# **Overview:**

This study focuses on finding an appropriate model for predicting the log solubility of differentcompounds based on their molecular weight and polar surface area using linear regression.

Predicting solubility of compounds in different aqueous solution can be beneficial for drugs preparation, chemical processing and several industrial applications.The main objective of the study is to find a predictive model for log solubility of compounds using machine learning algorithm, namely linear regression. Exploratory Data Analysis, data pre-processing and implementation different statistical models are used for comprehensive comparison.

The dataset used in this study contains molecular properties and corresponding solubility values. The primary variables include **Molecular Weight** and **Polar Surface Area**, key physicochemical properties of a compound, and **Measured Log Solubility (in mols/litre)**, the target variable for prediction.

On doing graphical data analysis on the predictor and Molecular weight an inverse relationship can be seen hence a linear regression is suitable for quantifying the trend.



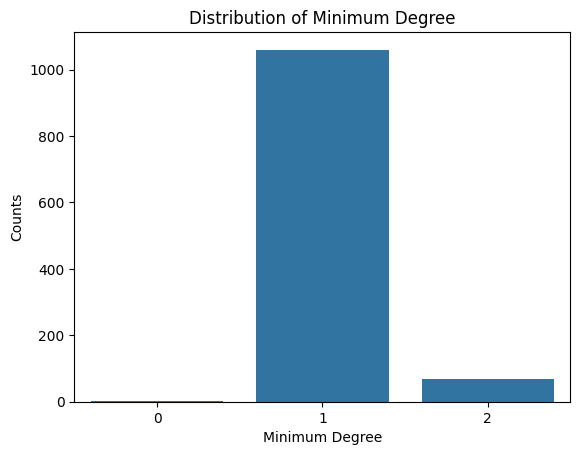
The dataset is then pre-processed to handle missing values, normalize features, and split into training and testing sets for evaluation with as much accuracy as possible. The accuracy of the training and testing dataset is then measured using measures like **Mean Square Error (MSE)** and **R2 score**.This research shows how data and machine learning can be used to solve real-world scientific problems efficiently.

# **Variable Description:**

The datasets contains the key variables which will help us to predict the log solubility of the compounds,

* Compound ID : Denoting the unique names of different compounds present in the dataset
* Minimum Degree: Minimum number of bonds any atom has.
* Molecular Weight: Total weight of the molecules,calculate based on the atoms present in the compounds.
* Number of H-Bond Donors: Number of groups in a molecule ready to donate a hydrogen bond e.g.: -NH and –OH
* Number of Rings: refers to the degree of unsaturation, which indicates the presence of rings and/or multiple bonds (double or triple) in a molecule
* Number of Rotatable Bonds: Number of rotatable singular bonds present.
* Polar Surface Area: Surface area covered by the polar atom such as oxygen or nitrogen.
* Measured Log Solubility: Logarithmic product of the solubility of compound in water.
* Simplified Molecular Input Line Entry System (SMILE): Molecular structure of different compounds.

# **Data Visualization**

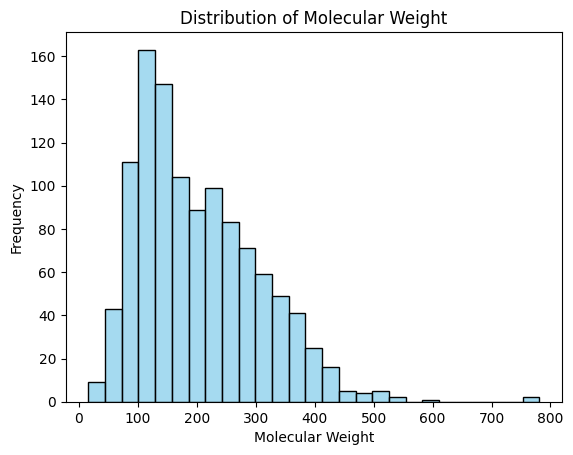
Minimum Degree:

|  |  |
| --- | --- |
| **Minimum Degree** | **Counts** |
| 0 | 1 |
| 1 | 1060 |
| 2 | 67 |

From the given graph we can distribution of minimum number of bonds a compound has.

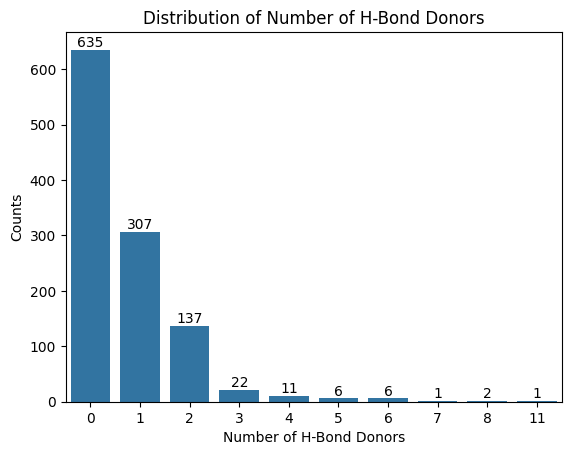
From the given graph we can clearly see that maximum compounds has at 1 bond and only 1 compound is devoid of any bonds.

Molecular Weight:



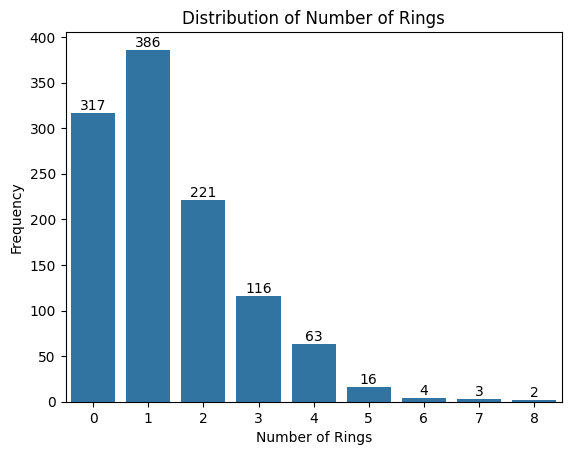
From the given graph it is evident that the molecular weight of molecules is positively skewed and the central tendency is around 100 with few outliers present after 500.

Number of H-bond donor:



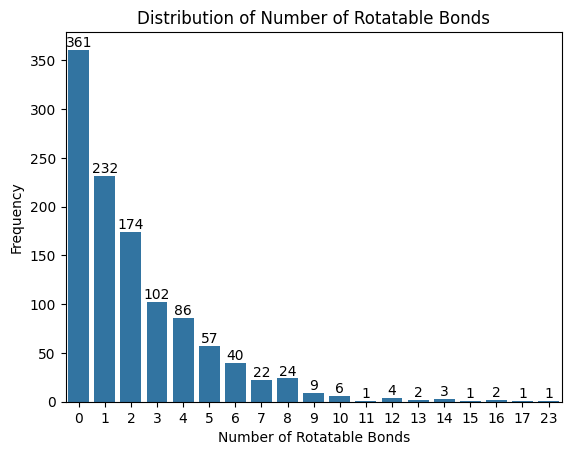
From the given graph it can be seen that maximum number of compounds don’t have any H-bond donating group and the graph is strictly decreasing indicating maximum compounds don’t have any H-bond donor.

Number of Rings:



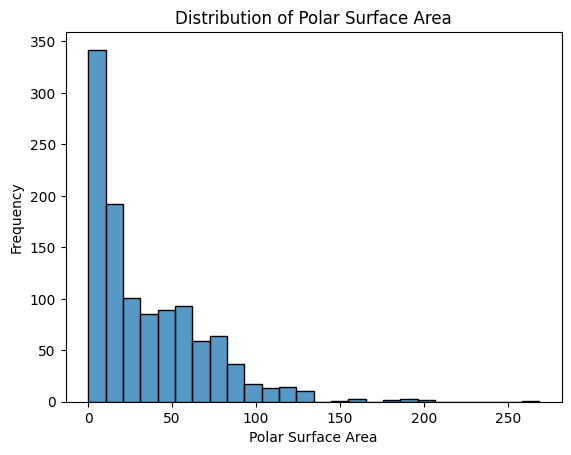
From the given graphs we can see the graph is highly positively skewed indicating that maximum compounds has only 1 ring and/or multiple bonds (double or triple) in a molecule.

Number of Rotatable Bonds:



The highest bar is at **0 rotatable bonds**, meaning most compounds do not have any rotatable bonds.As the number of rotatable bonds increases, the frequency decreases. For example, **1 rotatable bond** is found in 232 compounds, **2 bonds** in 174 compounds, and so on.Very few compounds have a high number of rotatable bonds. For example, only **1 compound** has 23 rotatable bonds. Indicating it is a decreasing in nature.

Polar Surface Area:



Most compounds have a **very small polar surface area**, with the highest number of compounds having values close to **zero**. Polar surface area increases as the number of compounds decreases.

# **Methodology:**

## **Data preparation:**

The data set is carefully checked for any missing values or null values, a thorough checking is done to check for presence of any null values or duplicated values in any of the rows.

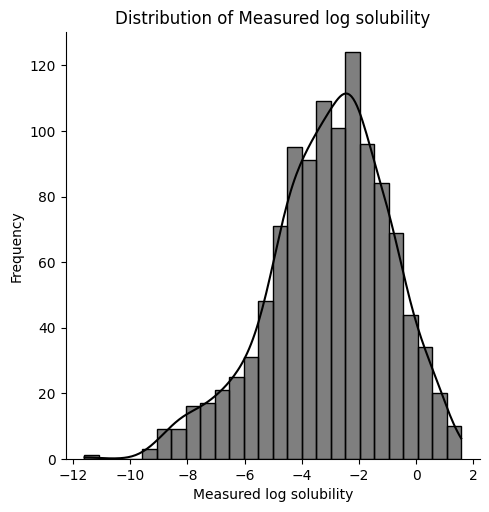
|  |  |  |
| --- | --- | --- |
| Variable Names | Null value present | Duplicate value present |
| Compound ID | 0 | 0 |
| Minimum Degree | 0 | 0 |
| Molecular Weight | 0 | 0 |
| Number of H-Bond Donors | 0 | 0 |
| Number of Rings | 0 | 0 |
| Number of Rotatable Bonds | 0 | 0 |
| Polar Surface Area | 0 | 0 |
| measured log solubility in (mol/ litre) | 0 | 0 |

After doing thorough checking of the dataset no missing or duplicated rows are found, confirming the data set is complete hence we can proceed with EDA without performing any operation on the dataset or hindrance.

## **Graphical Test for Normality of Dependent Variable:**

A Q-Q plot is used to graphically test for normality of the dependent variable i.e. “Measured Log Solubility”.

The distribution plot of dependent variable:

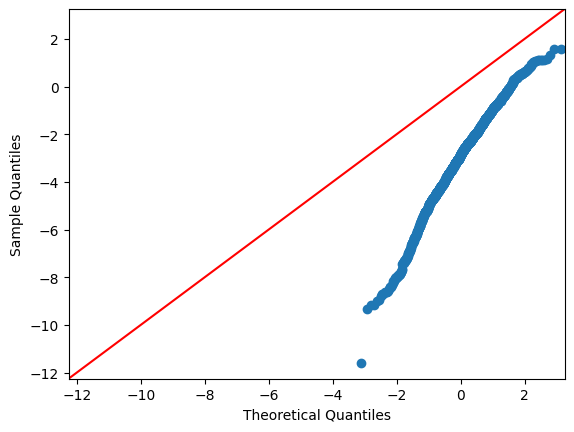


In the plot the **red line** represents the ideal normal distribution.

The actual points differs significantly from the ideal normal distribution specifically in the lower region. This indicates that there is presence of outliers in the dataset.

**Conclusion:** From the acquired plot we can conclude the variable **does not** follow normal distribution .i.e. the data is non-normal with negative skewed behaviour.

The Q-Q plot of the dependent variable



**Theoretical Test for Normality of Dependent Variable:**

Kolmogorov-Smirnov Test: Kolmogorov Smirnov test is a non parametric test used to determine if a given sample follows given distribution functions.

Let X1, X2, ....., Xn be a random sample of size n, where n=1128 and F(x) be the cumulative function of Gaussian distribution. Let Fn(x) be the empirical distribution function of the sample such that

Fn​(x) = (Number of observations ≤ x)/n​

Then to test:

H0: F(x)=Fn(x)

against

H1: Not H0

Let Dn be the test statistic such that,

Dn​= sup​∣Fn​(x) − F(x)∣

Let α be our level of significance such that α=5%.

Critical value is given as

Dn= c(α)/√n

c(α) is a constant based on the significance level α.

We can compute the **p-value**:

P(Dn>Dobs)=Qks((√n +0.12+0.11/√n)Dn).

Where,​ QKS​ is the **Kolmogorov-Smirnov survival function**.

Computation

* The observed value of our test statistic is coming as 0.04329377143271318 (~0.043)
* The p-value is coming as 0.02829143657915323 (~0.028)

Decision:

Since the p-value (0.028) is less than 0.05 (level of significance) we are rejecting H0 with 5% level of significance.

Conclusion:

The given data **does not** follow Normal distribution

Transformation to Gaussian distribution

We want to transform the given data to transform into a normal distribution; we will be using yeo-johnson transformation to achieve the desired result.

**Yeo-Johnson Transformation:**