In [ ]:
```
import pandas as pd
import numpy as np
```

Let´s create erk2 rotein inhibitors binary classification dataset for machine learning models.

We´ll count ***value <= 10000 nM as "Active" or "positive (1) class. **

And **value >=2000 nM as "Inactive" or "Negative (0)" class.**

Bioactivity data -> data analysis-> create binary classification dataset("Active/Inactve")

In [6]:
```
x=pd.read_csv("erk2.csv",sep=";")
```

In [34]:
```
x
```

Out[34]:

	Molecule ChEMBL ID	Molecule Name	Molecule Max Phase	Molecular Weight	#RO5 Violations	AlogP	Compound Key	
0	CHEMBL1351580	NaN	0	316.36	0	4.18	SID850528	
1	CHEMBL1493884	NaN	0	210.30	0	1.70	SID3712084	
2	CHEMBL1386667	NaN	0	460.56	0	1.63	SID7966145	
3	CHEMBL1455249	NaN	0	230.31	0	2.37	SID859084	
4	CHEMBL1451172	NaN	0	499.04	0	3.64	SID7967243	
...	...	...	...	...	...	...	...	
23225	CHEMBL4520788	NaN	0	448.59	1	5.58	None	CNC
23226	CHEMBL2180604	TAK-593	1	445.48	0	3.47	TAK-593	Cc1c
23227	CHEMBL4561806	NaN	0	242.30	0	2.62	None	
23228	CHEMBL3658647	NaN	0	477.55	0	3.36	BDBM157436	Cc1c
23229	CHEMBL3658846	NaN	0	338.37	0	2.24	BDBM157643	C

23230 rows × 45 columns

In [9]:
```
x["Standard Value"]
```

Out[9]:
```
0        25118.900
1        39810.700
2        31622.800
3        15848.900
4         7943.300
             ...
23225     4570.000
23226    30000.000
23227       89.000
23228        1.348
23229      376.200
Name: Standard Value, Length: 23230, dtype: float64
```

In [13]:
```
columns=pd.DataFrame(x.columns, columns=["column_name"])
```

In [15]:
```
columns
```

Out[15]:

	column_name
0	Molecule ChEMBL ID
1	Molecule Name
2	Molecule Max Phase
3	Molecular Weight
4	#RO5 Violations
5	AlogP
6	Compound Key
7	Smiles
8	Standard Type
9	Standard Relation
10	Standard Value

In [45]:
```
x1=x[["Molecule ChEMBL ID","Standard Value","Smiles","Standard Type","Standar
```

In [46]:
```
x1
```

Out[46]:

	Molecule ChEMBL ID	Standard Value	Smiles	Star
0	CHEMBL1351580	25118.900	COc1cccc(CSc2nnc(-c3ccc(F)cc3)o2)c1	Pc
1	CHEMBL1493884	39810.700	NC(=O)c1c(N)sc2c1CCCCC2	Pc
2	CHEMBL1386667	31622.800	COc1ccc(N2CCN(CCCNS(=O)(=O)c3ccc4c(c3)oc(=O)n4...	Pc
3	CHEMBL1455249	15848.900	CCCCc1nc2[nH]nc(N)c2c2c1CCC2	Pc
4	CHEMBL1451172	7943.300	O=C(C1CCN(S(=O)(=O)c2cccc3cccnc23)CC1)N1CCN(c2...	Pc
...	...	...	...	
23225	CHEMBL4520788	4570.000	CNCc1ccccc1-c1csc([C@H](C)Nc2nc(C)nc3cc(OC)c(O...	
23226	CHEMBL2180604	30000.000	Cc1cc(C(=O)Nc2cc(Oc3ccc4nc(NC(=O)C5CC5)cn4n3)c...	app

In [47]:
```
x1["Standard Type"].value_counts()
```

```
Activity                                  391
Residual Activity                         316
Kd apparent                               243
Kd                                        206
% Control                                  95
Ka                                         10
T1/2                                       10
Kdiss                                      10
Residual activity                           9
FC                                          4
% Ctrl                                      3
NT                                          3
Residual_activity                           3
EC50                                        2
Control                                     2
Inihibition                                 1
INH                                         1
Ratio IC50                                  1
% Residual activity with Skepinone-L        1
Name: Standard Type, dtype: int64
```

In [49]:
```
x2=x1[x1["Standard Units"].str.contains("nM", na=False)] #None missing data
```

In [50]:
```
x2
```

Out[50]:

	Molecule ChEMBL ID	Standard Value	Smiles	Standa Ty
0	CHEMBL1351580	25118.900	COc1cccc(CSc2nnc(-c3ccc(F)cc3)o2)c1	Poten
1	CHEMBL1493884	39810.700	NC(=O)c1c(N)sc2c1CCCCC2	Poten
2	CHEMBL1386667	31622.800	COc1ccc(N2CCN(CCCNS(=O)(=O)c3ccc4c(c3)oc(=O)n4...	Poten
3	CHEMBL1455249	15848.900	CCCCc1nc2[nH]nc(N)c2c2c1CCC2	Poten
4	CHEMBL1451172	7943.300	O=C(C1CCN(S(=O)(=O)c2cccc3cccnc23)CC1)N1CCN(c2...	Poten
...	...	...	...	
23224	CHEMBL4538174	91.000	C[C@@H](NC(=O)Nc1cc2[nH]ncc2c(CO)n1)c1ccccc1	IC
23225	CHEMBL4520788	4570.000	CNCc1ccccc1-c1csc([C@H](C)Nc2nc(C)nc3cc(OC)c(O...	IC
23226	CHEMBL2180604	30000.000	Cc1cc(C(=O)Nc2cc(Oc3ccc4nc(NC(=O)C5CC5)cn4n3)c...	appare
23228	CHEMBL3658647	1.348	Cc1cc(-c2n[nH]c3cc(NC(=O)NC4CCN(S(C)(=O)=O)c5c...	IC
23229	CHEMBL3658846	376.200	Cc1cc(-c2n[nH]c3cc(NC(=O)NCC4CCO4)ncc23)ccn1	IC

19068 rows × 6 columns

In [83]:
```
x2=x2.sort_values("Standard Value", ascending=True)
```

In [84]:
```
x2
```

Out[84]:

	Molecule ChEMBL ID	Standard Value	Smiles	Sta
22948	CHEMBL4115001	4.310000e-03	Nc1ncc([C@@H]2CC[C@@H](O)[C@H](O)C2)nc1-c1ccc(...	
57	CHEMBL4111166	5.000000e-03	NC[C@@H](NC(=O)c1ccc(-c2nc([C@@H]3CC[C@@H](O)[...	
22482	CHEMBL3904235	5.500000e-03	Nc1ncc([C@H]2CC[C@H](O)[C@@H](O)C2)nc1-c1ccc(C...	
20268	CHEMBL3980387	6.120000e-03	NC[C@@H](NC(=O)c1ccc(-c2nc(C3CCOCC3)cnc2N)cc1F...	
2071	CHEMBL4107592	6.650000e-03	CNC[C@@H](NC(=O)c1ccc(-c2nc([C@H]3CC[C@H](O)CC...	
...	...	...	...	
14970	CHEMBL4128535	1.000000e+06	CCOc1cc(N)cc(C(F)(F)F)c1	
16699	CHEMBL4129626	1.000000e+06	Nc1ccc2nc(-c3ccc(F)cc3)nn2c1	
15763	CHEMBL4128128	1.000000e+06	O=C1c2cccc(F)c2CN1[C@H]1CCCNC1	
894	CHEMBL2297162	1.000000e+06	Cc1nc[nH]c1[C@H]1c2nc[nH]c2CCN1Cc1nc2ccccc2[nH]1	
10678	CHEMBL1350100	5.011872e+06	CCC(=O)Nc1cc(C(=O)NCC2CCCN2CC)c(OC)cc1N(C)C	P

17815 rows × 6 columns

In [68]:
```
x2.head()
```

Out[68]:

	Molecule ChEMBL ID	Standard Value	Smiles	Standard Type	Standard Relation	Standard Units
22948	CHEMBL4115001	0.00431	Nc1ncc([C@@H]2CC[C@@H](O)[C@H](O)C2)nc1-c1ccc(...	IC50	'='	nM
57	CHEMBL4111166	0.00500	NC[C@@H](NC(=O)c1ccc(-c2nc([C@@H]3CC[C@@H](O)[...	IC50	'='	nM
22482	CHEMBL3904235	0.00550	Nc1ncc([C@H]2CC[C@H](O)[C@@H](O)C2)nc1-c1ccc(C...	IC50	'='	nM
20268	CHEMBL3980387	0.00612	NC[C@@H](NC(=O)c1ccc(-c2nc(C3CCOCC3)cnc2N)cc1F...	IC50	'='	nM
2071	CHEMBL4107592	0.00665	CNC[C@@H](NC(=O)c1ccc(-c2nc([C@H]3CC[C@H](O)CC...	IC50	'='	nM

In [69]:
```
x2.tail()
```

Out[69]:

	Molecule ChEMBL ID	Standard Value	Smiles	Stan
19198	CHEMBL4129032	1000000.0	O=C1c2ccc(F)cc2CN1C1CCNCC1	
12442	CHEMBL2297161	1000000.0	Cc1nc[nH]c1[C@@H]1c2nc[nH]c2CCN1Cc1nc2ccccc2[nH]1	
19020	CHEMBL4126333	1000000.0	O=C(O)c1ccc2nc(-c3ccccc3F)nn2c1	
20022	CHEMBL4127417	1000000.0	Fc1ccc2[nH]cc(C3CCCNC3)c2c1	
10678	CHEMBL1350100	5011872.3	CCC(=O)Nc1cc(C(=O)NCC2CCCN2CC)c(OC)cc1N(C)C	Pot

In [70]:
```
x2["Molecule ChEMBL ID"].value_counts()
```

Out[70]:
```
CHEMBL388978     12
CHEMBL3590107    12
CHEMBL4538174    10
CHEMBL3544964    10
CHEMBL3590106     9
                 ..
CHEMBL2004771     1
CHEMBL2000894     1
CHEMBL2003768     1
CHEMBL1993243     1
CHEMBL1350100     1
Name: Molecule ChEMBL ID, Length: 17815, dtype: int64
```

In [72]:
```
CHEML388978=x2[x2["Molecule ChEMBL ID"].str.contains("CHEMBL388978")]
```

In [73]:
```
CHEML388978
```

Out[73]:

	Molecule ChEMBL ID	Standard Value	Smiles	Standard Type	Standard Relation	Standard Units
21341	CHEMBL388978	1.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'<'	nM
17863	CHEMBL388978	2.5	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM
7769	CHEMBL388978	370.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM
6827	CHEMBL388978	370.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM
21376	CHEMBL388978	659.1	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM
12956	CHEMBL388978	1380.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM
11784	CHEMBL388978	3948.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM
6820	CHEMBL388978	4491.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM
17518	CHEMBL388978	7300.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	Kd	'='	nM
1689	CHEMBL388978	7300.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	Kd	'='	nM
21951	CHEMBL388978	8451.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM
19185	CHEMBL388978	34000.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM

```
Bioactivity data -> data analysis_> create binary classification
dataset("Active", "Inactive")
```

In [74]:
```
active=x2[x2["Standard Value"]<=1000]
inactive=x2[x2["Standard Value"]>20000]
```

In [75]:
```
d1=active[active["Molecule ChEMBL ID"].str.contains("CHEMBL388978")]
```

In [76]:
```
d1
```

Out[76]:

	Molecule ChEMBL ID	Standard Value	Smiles	Standard Type	Standard Relation	Standard Units
21341	CHEMBL388978	1.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'<'	nM
17863	CHEMBL388978	2.5	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM
7769	CHEMBL388978	370.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM
6827	CHEMBL388978	370.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM
21376	CHEMBL388978	659.1	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM

In [77]:
```
d2=inactive[inactive["Molecule ChEMBL ID"].str.contains("CHEMBL388978")]
```

In [78]:
```
d2
```

Out[78]:

	Molecule ChEMBL ID	Standard Value	Smiles	Standard Type	Standard Relation	Standard Units
19185	CHEMBL388978	34000.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM

In [85]:
```
x2.drop_duplicates("Molecule ChEMBL ID", inplace=True)
```

In [94]:
```
x2=x2.reset_index(drop=True)
```

In [97]:
```
x2=pd.DataFrame(x2)
```

In [98]:
```
x2
```

Out[98]:

	Molecule ChEMBL ID	Standard Value	Smiles	Sta
0	CHEMBL4115001	4.310000e-03	Nc1ncc([C@@H]2CC[C@@H](O)[C@H](O)C2)nc1-c1ccc(...	
1	CHEMBL4111166	5.000000e-03	NC[C@@H](NC(=O)c1ccc(-c2nc([C@@H]3CC[C@@H](O)[...	
2	CHEMBL3904235	5.500000e-03	Nc1ncc([C@H]2CC[C@H](O)[C@@H](O)C2)nc1-c1ccc(C...	
3	CHEMBL3980387	6.120000e-03	NC[C@@H](NC(=O)c1ccc(-c2nc(C3CCOCC3)cnc2N)cc1F...	
4	CHEMBL4107592	6.650000e-03	CNC[C@@H](NC(=O)c1ccc(-c2nc([C@H]3CC[C@H](O)CC...	
...	...	...	...	
17810	CHEMBL4128535	1.000000e+06	CCOc1cc(N)cc(C(F)(F)F)c1	
17811	CHEMBL4129626	1.000000e+06	Nc1ccc2nc(-c3ccc(F)cc3)nn2c1	
17812	CHEMBL4128128	1.000000e+06	O=C1c2cccc(F)c2CN1[C@H]1CCCNC1	
17813	CHEMBL2297162	1.000000e+06	Cc1nc[nH]c1[C@H]1c2nc[nH]c2CCN1Cc1nc2ccccc2[nH]1	
17814	CHEMBL1350100	5.011872e+06	CCC(=O)Nc1cc(C(=O)NCC2CCCN2CC)c(OC)cc1N(C)C	P

17815 rows × 6 columns

In [109]:
```
columns
```

Out[109]:

	column_name
0	Molecule ChEMBL ID
1	Molecule Name
2	Molecule Max Phase
3	Molecular Weight
4	#RO5 Violations
5	AlogP
6	Compound Key
7	Smiles
8	Standard Type
9	Standard Relation
10	Standard Value
11	Standard Units
12	pChEMBL Value
13	Data Validity Comment
14	Comment
15	Uo Units
16	Ligand Efficiency BEI
17	Ligand Efficiency LE
18	Ligand Efficiency LLE
19	Ligand Efficiency SEI
20	Potential Duplicate
21	Assay ChEMBL ID
22	Assay Description
23	Assay Type
24	BAO Format ID
25	BAO Label
26	Assay Organism
27	Assay Tissue ChEMBL ID
28	Assay Tissue Name
29	Assay Cell Type
30	Assay Subcellular Fraction
31	Assay Parameters
32	Assay Variant Accession
33	Assay Variant Mutation

	column_name
34	Target ChEMBL ID
35	Target Name
36	Target Organism
37	Target Type
38	Document ChEMBL ID
39	Source ID
40	Source Description
41	Document Journal
42	Document Year
43	Cell ChEMBL ID
44	Properties

In [110]:
```
columns.to.csv("erk2_columns.csv",sep=")
```

```
  File "C:\Users\L03121~1\AppData\Local\Temp/ipykernel_22444/2029187873.py"
, line 1
    columns.to.csv("erk2_columns.csv",sep=")
                                          ^
SyntaxError: EOL while scanning string literal
```

In [111]:
```
x.head()
```

Out[111]:

	Molecule ChEMBL ID	Molecule Name	Molecule Max Phase	Molecular Weight	#RO5 Violations	AlogP	Compound Key	
0	CHEMBL1351580	NaN	0	316.36	0	4.18	SID850528	
1	CHEMBL1493884	NaN	0	210.30	0	1.70	SID3712084	N(
2	CHEMBL1386667	NaN	0	460.56	0	1.63	SID7966145	CO
3	CHEMBL1455249	NaN	0	230.31	0	2.37	SID859084	CCCCc
4	CHEMBL1451172	NaN	0	499.04	0	3.64	SID7967243	(=O)c2cccc3

5 rows × 45 columns

In [100]:
```
x2["Molecule ChEMBL ID"].value_counts()
```

Out[100]:
```
CHEMBL4115001    1
CHEMBL1447777    1
CHEMBL1440725    1
CHEMBL1426223    1
CHEMBL1582279    1
                ..
CHEMBL2004716    1
CHEMBL1562756    1
CHEMBL336961     1
CHEMBL1982660    1
CHEMBL1350100    1
Name: Molecule ChEMBL ID, Length: 17815, dtype: int64
```

In [103]:
```
CHEML388978
```

Out[103]:

	Molecule ChEMBL ID	Standard Value	Smiles	Standard Type	Standard Relation	Standard Units
21341	CHEMBL388978	1.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'<'	nM
17863	CHEMBL388978	2.5	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM
7769	CHEMBL388978	370.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM
6827	CHEMBL388978	370.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM
21376	CHEMBL388978	659.1	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM
12956	CHEMBL388978	1380.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM
11784	CHEMBL388978	3948.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM
6820	CHEMBL388978	4491.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM
17518	CHEMBL388978	7300.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	Kd	'='	nM
1689	CHEMBL388978	7300.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	Kd	'='	nM
21951	CHEMBL388978	8451.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM
19185	CHEMBL388978	34000.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	nM

In [106]:
```
CHEML388978["Standard Value"].mean()
```

Out[106]:
```
5689.383333333334
```

In [107]:
```
CHEML388978["new value"]=CHEML388978["Standard Value"].mean()
```

```
C:\Users\L03121898\AppData\Local\Schrodinger\PyMOL2\envs\Cadi22\lib\site-pa
ckages\ipykernel_launcher.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://p
andas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-vi
ew-versus-a-copy)
  """Entry point for launching an IPython kernel.
```

In [108]:
```
CHEML388978
```

Out[108]:

	Molecule ChEMBL ID	Standard Value	Smiles	Standard Type	Standard Relation	Standa Un
21341	CHEMBL388978	1.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'<'	r
17863	CHEMBL388978	2.5	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	r
7769	CHEMBL388978	370.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	r
6827	CHEMBL388978	370.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	r
21376	CHEMBL388978	659.1	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	r
12956	CHEMBL388978	1380.0	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3...	IC50	'='	r
11784	CHEMBL388978	3948.0	CN[C@@H]1C[C@H]2O[C@@]	IC50	'='	r