Group Members

Andres C0931978

Johan C0931102

Amandeep C0937432

Brian C0940439

Manjot C0938094

Anju C0935847

Sarita C0938516

Amita C0935607

Load Dataset in Chunks Loading the file in chunks reduces memory usage and speeds up the process for large files.

In [2]:
```
file_path = r"C:\Users\USUARIO\Desktop\projectAD\PubChem_compound_cache_p-AA8AaJ

# Load in chunks
chunk_size = 10000  # Define chunk size
chunks = []

for chunk in pd.read_csv(file_path, chunksize=chunk_size, low_memory=False):
    chunks.append(chunk)

# Combine chunks into a single DataFrame
data = pd.concat(chunks, ignore_index=True)
```

Check Encoding If the file has special characters, ensure the correct encoding is used. Common encodings include utf-8, latin1, etc.

In [3]:
```
data = pd.read_csv(file_path, encoding='utf-8')
```

Optimize Data Types You can specify data types for columns while loading to reduce memory usage:

In [4]:
```
data = pd.read_csv(
    file_path,
    dtype={
        "mw": "float32",  # Example for numeric columns
        "mf": "string",   # Example for string columns
    },
    low_memory=False
)
```

Debug File Structure Print the first few lines of the file directly to check if it's accessible:

In [5]:
```
with open(file_path, 'r', encoding='utf-8') as file:
    for _ in range(5):  # Print the first 5 lines
        print(file.readline())
```

cid,cmpdname,cmpdsynonym,mw,mf,polararea,complexity,xlogp,heavycnt,hbonddonor,hbondacc,rotbonds,inchi,isosmiles,canonicalsmiles,inchikey,iupacname,exactmass,monoisotopicmass,charge,covalentunitcnt,isotopeatomcnt,totalatomstereocnt,definedatomstereocnt,undefinedatomstereocnt,totalbondstereocnt,definedbondstereocnt,undefinedbondstereocnt,pclidcnt,gpidcnt,gpfamilycnt,meshheadings,annothits,annothitcnt,aids,cidcdate,sidsrcname,depcatg,annotation

1,Acetyl-DL-carnitine,"Acetyl-DL-carnitine|acetylcarnitine|14992-62-2|DL-Acetylcarnitine|DL-O-Acetylcarnitine|3-acetyloxy-4-(trimethylazaniumyl)butanoate|Acetylcarnitine|UNII-07OP6H4V4A|3-(acetyloxy)-4-(trimethylazaniumyl)butanoate|07OP6H4V4A|1-Propanaminium, 2-(acetyloxy)-3-carboxy-N,N,N-trimethyl-, inner salt|ACETYLCARNITINE, DL-|ACETYLCARNITINE DL-FORM|Ammonium, (3-carboxy-2-hydroxypropyl)trimethyl-, hydroxide, inner salt, acetate, DL-|870-77-9|DTXSID2048117|ACETYLCARNITINE, (+/-)-|ACETYLCARNITINE RACEMATE [MI]|3-(acetyloxy)-4-(trimethylammonio)butanoate|1-Propanaminium, 2-(acetoxy)-3-carboxy-N,N,N-trimethyl-, hydroxide, inner salt, (+-)-|(3-CARBOXY-2-HYDROXYPROPYL)TRIMETHYLAMMONIUM HYDROXIDE INNER SALT ACETATE|AMMONIUM, (3-CARBOXY-2-HYDROXYPROPYL)TRIMETHYL-, HYDROXIDE, INNER SALT, ACETATE|CHEBI:73024|(+/-)-acetylcarnitine|bmse000142|SCHEMBL69781|O-ACETYL-DL-CARNITINE|D/L-ACETYL-CARNITINE|ACETYLCARNITINE RACEMATE|ACETYL CARNITINE [INCI]|DTXCID5028089|LMFA07070060|3-Acetoxy-4-(trimethylammonio)butanoate|AS-82719|HY-126358|CS-0102945|NS00074255|G90449|Q27140241|2-(ACETYLOXY)-3-CARBOXY-N,N,N-TRIMETHYL-1-PROPANAMINIUM INNER SALT|1-Propanaminium, 2-(acetoxy)-3-carboxy-N,N,N-trimethyl-, hydroxide, inner salt, (+-)- (9CI)|1-Propanaminium, 2-(acetoxy)-3-carboxy-N,N,N-trimethyl-, hydroxide, inner salt, (+-)-(9CI)",203.240,C9H17NO4,66.400,214.000,0.400,14,0,4,5,"InChI=1S/C9H17NO4/c1-7(11)14-8(5-9(12)13)6-10(2,3)4/h8H,5-6H2,1-4H3",CC(=O)OC(CC(=O)[O-])C[N+](C)(C)C,CC(=O)OC(CC(=O)[O-])C[N+](C)(C)C,RDHQFKQIGNGIED-UHFFFAOYSA-N,3-acetyloxy-4-(trimethylazaniumyl)butanoate,203.116,203.116,0,1,0,1,0,1,0,0,0,2784,3332,1232,Acetylcarnitine,Interactions and Pathways|Chemical and Physical Properties|Classification|Drug and Medication Information|Literature|Patents|Pharmacology and Biochemistry|Use and Manufacturing|Spectral Information|Taxonomy,10,"",20050623,"10X CHEM|3WAY PHARM INC|A2B Chem|AA BLOCKS|AbaChemScene|ABI Chem|AHH Chemical co.,ltd|Alfa Chemistry|Amadis Chemical|Aribo Reagent|AstaTech, Inc.|BenchChem|Biological Magnetic Resonance Data Bank (BMRB)|BLD Pharm|BOC Sciences|ChEBI|ChemIDplus|Chemieliva Pharmaceutical Co., Ltd|CHIRALEN|Clearsynth|Collaborative Drug Discovery, Inc.|DiscoveryGate|Egon Willighagen, Department of Bioinformatics - BiGCaT, Maastricht University|eMolecules|EPA DSSTox|FDA Global Substance Registration System (GSRS)|Finetech Industry Limited|Google Patents|Hairui Chemical|Human Metabolome Database (HMDB)|iChemical Technology USA Inc|InvivoChem|Key Organics/BIONET|LeadScope|LIPID MAPS|MassBank of North America (MoNA)|Mcule|MedChemexpress MCE|Metabolomics Workbench|MolGenie|MolPort|NextBio|NORMAN Suspect List Exchange|NovoSeek|Parchem|PATENTSCOPE (WIPO)|Probes & Drugs portal|PubChem Reference Collection|RR Scientific|SCRIPDB|Smolecule|Springer Nature|SureChEMBL|THE BioTek|Thomson Pharma|ToxPlanet|Wikidata|WikiPathways|ZeroPM - Zero Pollution of Persistent, Mobile Substances",Chemical Vendors|Curation Efforts|Governmental Organizations|Journal Publishers|Legacy Depositors|NIH Initiatives|Research and Development|Subscription Services,"D002491 - Central Nervous System Agents > D018697 - Nootropic Agents|D018977 - Micronutrients > D014815 - Vitamins|N - Nervous system > N06 - Psychoanaleptics > N06B - Psychostimulants, agents used for adhd and nootropics"

6,Dinitrochlorobenzene,"1-chloro-2,4-dinitrobenzene|2,4-Dinitrochlorobenzene|97-00-7|Dinitrochlorobenzene|DNCB|Chlorodinitrobenzene|CDNB|Benzene, 1-chloro-2,4-dinitro-|4-Chloro-1,3-dinitrobenzene|2,4-Dinitrophenyl chloride|Dinitrochlorobenzol|6-Chloro-1,3-dinitrobenzene|1,3-Dinitro-4-chlorobenzene|2,4-Dinitro-1-chlorobenzene|1-Chloro-2,4-dinitrobenzol|ClDNB|DNPCl|1-Chlor-2,4-dinitrobenzene|1-Cloro-2,4-dinitrobenzene|Caswell No. 389C|1-Chloor-2,4-dinitrobenzeen|Dinitrochlorobenzene (VAN)|1-Chlor-2,4-dinitrobenzol|NSC 6292|1-chloro-2,4-dinitro-benzene|CCRIS 1799|UNII-GE3IBT7BMN|CHEBI:34718|HSDB 5306|NSC-6292|EINECS 202-551-4|1-CHLORO-2,4-DINITROBENZENE-3,5,6-D3|1-Chloor-2,4-dinitrobenzeen [Dutch]|1-Chloro-2,4-dinitrobenzeen [Dutch]|1-Chloro-2,4-dinitrobenzeen|EPA Pesticide Chemical Code 055102|1-Cloro

-2,4-dinitrobenzene [Italian]|347840-12-4|DTXSID6020278|AI3-01053|NSC6292|GE3IBT7BMN|DTXCID00278|1-chloro-2,4-dinitro benzene|EC 202-551-4|MFCD00007075|2,4-DINITROCHLOROBENZENE [MART.]|1-CHLORO-2,4-DINITROBENZENE [MI]|1-CHLORO-2,4-DINITROBENZENE [HSDB]|1-CHLORO-2,4-DINITROBENZENE [WHO-DD]|CAS-97-00-7|2,4-DINITROCHLOROBENZENE (MART.)|SMR000857169|1-CHLORO-2,4-DINITROBENZEEN (DUTCH)|1-CLORO-2,4-DINITROBENZENE (ITALIAN)|2,4-dinitro chlorobenzene|chloro-2,4-dinitrobenzene|1-Chloro-2,4-dinitrobenzol [German]|1-CHLORO-2, 4-DINITROBENZENE (DINITROCHLOROBENZENE)|1-CHLORO-2, 4-DINITROBENZENE {DINITROCHLOROBENZENE}|2,4dinitrochlorobenzene|1Chlor2,4dinitrobenzol|2,4-dinitrochlorbenzene|1Chlor2,4dinitrobenzene|1Chloro2,4dinitrobenzol|1Cloro2,4dinitrobenzene|2-4 dinitrochlorobenzene|1,3Dinitro4chlorobenzene|1Chloor2,4dinitrobenzeen|1Chloro2,4dinitrobenzeen|2,4Dinitro1chlorobenzene|4Chloro1,3dinitrobenzene|6Chloro1,3dinitrobenzene|WLN: WNR BG ENW|2,4Dinitrophenyl chloride|2,4-dinitro-chlorobenzene|Epitope ID:110163|2,4-dinitro-chloro-benzene|2,4dinitro-1-chlorobenzene|Benzene, 1chloro2,4dinitro|SCHEMBL39251|1-CHLORODINITROBENZENE|2-chloro-1,5-dinitrobenzene|MLS001332459|MLS001332460|BIDD:ER0694|CHEMBL292687|HMS2233O04|BCP27853|STR01511|Tox21_201956|Tox21_302802|BBL009322|BDBM50458521|STK387094|1-Chloro-2,4-dinitrobenzene, 97%|AKOS000118946|1-Chloro-2,4-dinitrobenzene, ~95%|DB11831|1-Chloro-2,4-dinitrobenzene, >=99%|NCGC00164061-01|NCGC00164061-02|NCGC00164061-03|NCGC00256396-01|NCGC00259505-01|DB-057658|NS00003347|EN300-18084|Q209216|1-Chloro-2,4-dinitrobenzene, technical grade, 95%|W-100123|Z57160126|1-Chloro-2,4-dinitrobenzene 100 microg/mL in Methanol|3-(3-hydroxy-2-methyl-4-oxo-1-pyridyl)propanoic acid|F1908-0126",202.550,C6H3ClN2O4,91.600,224.000,2.300,13,0,4,0,InChI=1S/C6H3ClN2O4/c7-5-2-1-4(8(10)11)3-6(5)9(12)13/h1-3H,C1=CC(=C(C=C1[N+](=O)[O-])[N+](=O)[O-])Cl,C1=CC(=C(C=C1[N+](=O)[O-])[N+](=O)[O-])Cl,VYZAHLCBVHPDDF-UHFFFAOYSA-N,"1-chloro-2,4-dinitrobenzene",201.978,201.978,0,1,0,0,0,0,0,0,0,6887,12139,4796,Dinitrochlorobenzene,Biological Test Results|Interactions and Pathways|Chemical and Physical Properties|Classification|Drug and Medication Information|Food Additives and Ingredients|Identification|Literature|Patents|Pharmacology and Biochemistry|Safety and Hazards|Toxicity|Use and Manufacturing|Associated Disorders and Diseases|Spectral Information|Biological Test Results: Active|Biological Test Results: Micromolar,17,155|157|161|165|167|175|179|192|220|300|302|1189|1199|1205|1208|1469|1479|1850|1863|1899|1903|1906|1947|1950|1962|1974|1987|1996|2016|2023|2025|2029|2052|2057|2066|2101|2129|2130|2174|2177|2234|2235|2280|2300|2380|2391|2435|2445|2462|2517|2520|2521|2524|2540|2544|2557|2599|2606|2629|2650|2661|2676|2685|2690|2716|2717|2718|2751|2796|2797|2805|2806|2825|23442|78385|78386|78387|78519|78520|78521|78522|78523|78534|78535|200690|267573|434955|434962|434973|434989|435003|435005|435022|435030|449728|449748|449762|449763|463074|463079|463082|463104|463141|463190|463195|463210|463212|463213|463254|485270|485272|485273|485275|485294|485297|485298|485313|485317|485341|485344|485346|485347|485349|485358|485360|485364|485367|488837|488839|488847|488862|488890|488895|488896|488899|488965|488966|488975|488977|489028|489030|489031|492947|492953|492956|492972|493005|493008|493011|493012|493014|493036|493056|493084|493087|493091|493098|493131|493160|493187|493244|504326|504327|504329|504332|504333|504339|504357|504406|504411|504414|504423|504441|504444|504454|504462|504466|504467|504490|504523|504558|504577|504582|504621|504634|504648|504651|504652|504660|504690|504692|504700|504706|504707|504720|504734|504766|504775|504803|504810|504812|504832|504834|504842|504845|504847|504884|504891|504894|504937|540253|540263|540267|540275|540277|540295|540303|540308|540317|540336|540364|588334|588335|588342|588350|588351|588352|588354|588358|588391|588405|588413|588436|588453|588456|588458|588473|588475|588478|588489|588492|588493|588497|588499|588501|588511|588513|588514|588515|588516|588526|588527|588532|588533|588534|588535|588536|588537|588541|588543|588544|588545|588546|588547|588549|588579|588590|588591|588621|588627|588664|588674|588675|588676|588689|588692|588726|588727|588795|588814|588819|588850|588852|588855|588856|589004|589008|589009|589017|589018|589019|602123|602141|602162|602163|602179|602229|602233|602244|602247|602248|602250|602252|602261|602274|602281|602310|602313|602329|602332|602340|602342|602346|602363|602393|602396|602399|602405|602410|602429|602438|602440|602449|602481|623877|623901|624037|624038|624040|624125|624126|624127|624168|624169|624170|624171|624172|624173|624178|624202|624204|624246|624256|624263|624267|624268|624288|624296|624297|624304|624330|624352|624354|624377|624414|624415|624416|624417|624418|624463|624464|624465|624466|624467|624483|643387|643449|65

1550|651560|651572|651582|651596|651602|651610|651631|651632|651633|651634|651635
|651636|651640|651644|651647|651654|651658|651661|651687|651699|651702|651704|651
710|651711|651718|651723|651724|651725|651741|651743|651749|651751|651768|651777|
651778|651800|651819|651820|651821|651838|651957|651958|651965|651999|652010|6520
17|652025|652039|652048|652051|652054|652067|652104|652105|652106|652115|652126|6
52154|652162|652163|652197|652257|686940|686964|686970|686971|686978|686979|68699
2|686996|687014|687016|720504|720508|720509|720511|720516|720542|720543|720551|72
0552|720553|720579|720580|720582|720596|720634|720635|720636|720637|720647|720648
|720653|720659|720674|720675|720678|720679|720680|720681|720682|720683|720684|720
685|720686|720687|720691|720692|720693|720700|720702|720704|720706|720707|720708|
720709|720711|720719|720725|743012|743014|743015|743033|743035|743036|743040|7430
41|743042|743053|743054|743063|743064|743065|743066|743067|743069|743074|743075|7
43077|743078|743079|743080|743081|743083|743084|743085|743086|743091|743094|74312
2|743126|743139|743140|743191|743194|743199|743202|743203|743209|743210|743211|74
3212|743213|743215|743217|743218|743219|743220|743221|743222|743223|743224|743225
|743226|743227|743228|743238|743239|743240|743241|743242|743255|743266|743269|743
279|743288|743292|743397|743398|1053197|1159509|1159515|1159516|1159517|1159518|1
159519|1159520|1159521|1159523|1159524|1159525|1159526|1159527|1159528|1159529|11
59531|1159551|1159552|1159553|1159555|1159606|1224834|1224835|1224836|1224837|122
4838|1224839|1224840|1224841|1224842|1224843|1224844|1224845|1224846|1224847|1224
848|1224849|1224865|1224867|1224868|1224869|1224870|1224871|1224872|1224873|12248
74|1224875|1224876|1224877|1224878|1224879|1224880|1224881|1224882|1224883|122488
4|1224885|1224886|1224887|1224888|1224889|1224890|1224892|1224893|1224894|1224895
|1224896|1259241|1259242|1259243|1259244|1259247|1259248|1259313|1259318|1259364|
1259365|1259366|1259367|1259368|1259369|1259377|1259378|1259379|1259380|1259381|1
259382|1259383|1259384|1259385|1259386|1259387|1259388|1259390|1259391|1259392|12
59393|1259394|1259395|1259396|1259401|1259402|1259403|1259404|1259407|1259415|127
2365|1346378|1346784|1346795|1346798|1346799|1346824|1346829|1346859|1346877|1346
891|1346924|1346977|1346978|1346979|1346980|1346981|1346982|1347030|1347031|13470
32|1347033|1347034|1347035|1347036|1347037|1347038|1347041|1347056|1347071|134707
5|1347076|1347120|1347131|1347395|1347397|1347398|1347399|1384567|1388635|1409598
|1640020|1645758|1671190|1671196|1671197|1671198|1671199|1671200|1671201|1671463|
1671498|1745844|1745845|1749710|1749711|1794731|1794732|1794733|1794735|1794736|1
794738|1794739|1794740|1794742|1794745|1794746|1794748|1794751|1794752|1794753|17
94754|1794755|1794756|1794757|1794758|1794759|1794760|1794761|1794763|1794764|179
4765|1794766|1794767|1794768|1794769|1794771|1794772|1794773|1794774|1794775|1794
776|1794777|1794778|1794779|1794780|1794781|1794782|1794783|1794784|1794785|17947
86|1794787|1794788|1794789|1794790|1794792|1794793|1794794|1794795|1794796|179479
8|1794799|1794800|1919968|1919969|1919970|1920062|1920063|1920064|1920065|1920067
|1920068|1963577|1963578|1963579|1963580|1963581|1963582|1963583|1963584|1963585,
20050326,"10X CHEM|3B Scientific (Wuhan) Corp|3WAY PHARM INC|A&J Pharmtech CO., L
TD.|A2B Chem|AA BLOCKS|AAA Chemistry|abcr GmbH|ABI Chem|Acadechem|Achemica|Achemo
Scientific Limited|Achemtek|Acorn PharmaTech Product List|AHH Chemical co.,ltd|AK
os Consulting & Solutions|Alfa Chemistry|Amadis Chemical|Ambinter|AN PharmaTech|A
ngene Chemical|Anward|Apexmol|Aronis|Aurora Fine Chemicals LLC|Aurum Pharmatech L
LC|AZEPINE|BenchChem|Bic Biotech|BIDD|BindingDB|BioChemPartner|BioCyc|Biosynth|BO
C Sciences|Boerchem|Carcinogenic Potency Database (CPDB)|ChEBI|Chem-Space.com Dat
abase|Chembase.cn|ChEMBL|ChemDB|ChemExper Chemical Directory|Chemhere|Chemical Ca
rcinogenesis Research Information System (CCRIS)|ChemIDplus|Chemieliva Pharmaceut
ical Co., Ltd|ChemMol|ChemShuttle|ChemSpider|ChemTik|Collaborative Drug Discover
y, Inc.|Comparative Toxicogenomics Database (CTD)|CymitQuimica|Davey Lab, Departm
ent of Microbiology, NEIDL, Boston University|Debye Scientific Co., Ltd|Discovery
Gate|DrugBank|DSL Chemicals|DTP/NCI|Egon Willighagen, Department of Bioinformatic
s - BiGCaT, Maastricht University|eMolecules|Enamine|EPA DSSTox|EPA Substance Reg
istry Services|FDA Global Substance Registration System (GSRS)|Google Patents|Hai
rui Chemical|Human Metabolome Database (HMDB)|IBM|ICCB-Longwood Screening Facilit
y, Harvard Medical School|iChemical Technology USA Inc|Immune Epitope Database (I
EDB)|Innovapharm|ISpharm|IUPAC Digitized pKa Dataset|J&H Chemical Co.,ltd|Japan C
hemical Substance Dictionary (Nikkaji)|KCS Online|KEGG|Key Organics/BIONET|King S
cientific|labseeker|LeadScope|LGC Standards|Life Chemicals|MassBank Europe|MassBa

nk of North America (MoNA)|Matrix Scientific|Mcule|Metabolomics Workbench|MLSMR|MolCore|Molecule Market|Molepedia|MolGenie|MolPort|MP Biomedicals|MuseChem|National Center for Advancing Translational Sciences (NCATS)|Nature Chemistry|NextBio|NextMove Software|NIAID ChemDB|NIST Chemistry WebBook|NIST Mass Spectrometry Data Center|NMRShiftDB|NORMAN Suspect List Exchange|NovoSeek|Oakwood Products|OtavaChemicals|Parchem|PATENTSCOPE (WIPO)|Probes & Drugs portal|PubChem Reference Collection|RR Scientific|Santa Cruz Biotechnology, Inc.|SCRIPDB|Sigma-Aldrich|Sinfoo Biotech|Small Molecule Screening Facility, UW Madison|Smolecule|Springer Nature|SpringerMaterials|Starshine Chemical|SureChEMBL|SynQuest Laboratories|TCI (Tokyo Chemical Industry)|THE BioTek|Therapeutic Target Database (TTD)|Thermo Fisher Scientific|Thieme Chemistry|Thomson Pharma|TimTec|Tox21|ToxPlanet|Tractus|Vitas-M Laboratory|VladaChem|Wikidata|WikiPathways|Wiley|Win-Win Chemical|Wubei-Biochem|Wutech|Yick-Vic Chemicals & Pharmaceuticals (HK) Ltd.|Yuhao Chemical|ZeroPM - Zero Pollution of Persistent, Mobile Substances|ZINC|Zjartschem",Chemical Vendors|Curation Efforts|Governmental Organizations|Journal Publishers|Legacy Depositors|NIH Initiatives|Research and Development|Subscription Services,C308 - Immunotherapeutic Agent > C2139 - Immunostimulant|D009676 - Noxae > D007509 - Irritants|D019995 - Laboratory Chemicals > D007202 - Indicators and Reagents

11,"1,2-Dichloroethane","1,2-dichloroethane|Ethylene dichloride|107-06-2|Ethylene chloride|Ethane, 1,2-dichloro-|Dichloroethylene|Glycol dichloride|Dutch liquid|Ethane dichloride|Aethylenchlorid|Dichloro-1,2-ethane|sym-Dichloroethane|Dichloremulsion|1,2-DCE|1,2-Dichlorethane|Brocide|1,2-Bichloroethane|Dichlor-Mulsion|Bichlorure D'ethylene|Borer sol|Di-chlor-mulsion|Freon 150|alpha,beta-Dichloroethane|EDC (halocarbon)|1,2-Ethylene dichloride|Destruxol borer-sol|1,2-Ethylidene dichloride|Caswell No. 440|Aethylendichlorid|s-Dichloroethane|Ethyleendichloride|1,2-Dichlor-aethan [German]|Cloruro di ethene|Rcra waste number U077|1,2-Dicloroetano|RY Dichloro-1,2-ethane|1,2-Dichloraethan|Chlorure D'ethylene|1,2-Dichloorethaan|DCE|1,2-Dichlor-aethan|HSDB 65|CCRIS 225|1,2-Dichloro-Ethane|Ethylene dichloride [BSI:ISO]|HCC 150|ethylenedichloride|NCI-C00511|1,2dichloroethane|Dichlorure d'ethylene [ISO-French]|EDC|EINECS 203-458-1|ENT 1,656|MFCD00000963|1.2-dichloroethane|.alpha.,.beta.-Dichloroethane|CH2ClCH2Cl|DTXSID6020438|UNII-55163IJI47|AI3-01656|DTXCID40438|52399-93-6|ETHYLENE DICHLORIDE [MI]|CHEBI:27789|ETHYLENE DICHLORIDE [FCC]|ETHYLENE DICHLORIDE [ISO]|EC 203-458-1|ETHYLENE DICHLORIDE [HSDB]|1,2-DICHLOROETHANE [IARC]|ETHYLENE DICHLORIDE [MART.]|ETHYLENE DICHLORIDE [WHO-DD]|1,2-DICHLOROETHANE [USP-RS]|55163IJI47|UN 1184|ETHAMBUTOL HYDROCHLORIDE IMPURITY D [EP IMPURITY]|1, 2-dichloroethane|1,2-DICHLOROETHANE (IARC)|ETHYLENE DICHLORIDE (MART.)|Aethylenchlorid [German]|Dichlorure d'ethylene (ISO-French)|1,2-DICHLOROETHANE (USP-RS)|Ethyleendichloride [Dutch]|1,2-Dichloroethane, analytical standard|Cloruro di ethene [Italian]|1,2-Dichloorethaan [Dutch]|1,2-Dicloroetano [Italian]|Chlorure d'ethylene [French]|1,2 dichloroethane|Bichlorure d'ethylene [French]|Dichlorure d'ethylene|Dichloro-1,2-ethane [French]|ClCH2CH2Cl|1,2-Dichloroethane 100 microg/mL in Methanol|1, 2-DICHLOROETHANE (ETHYLENE DICHLORIDE)|1, 2-DICHLOROETHANE {ETHYLENE DICHLORIDE}|alpha, beta-dichloroethane|UN1184|RCRA waste no. U077|EPA Pesticide Chemical Code 042003|ETHAMBUTOL HYDROCHLORIDE IMPURITY D (EP IMPURITY)|ethylenechloride|-Dichloroethane|ethylendichloride|alpha,Bet|dichloro ethylene|ethylene-chloride|ethylene dichoride|1,2-DICHLOROETHANE, ACS|1,2dichlorethane|ehtylene dichloride|a,b-Dichloroethane|Ethene, dichloro-|1,2 dichlorethane|1,2 dichoroethane|1,2-dichloroetane|1,2-dichloroethan|1,2-dichoroethane|1,2-dicloroethane|1,2-dichioroethane|Ethylene, dichloro-|1,2-dichloroetharie|1 ,2-dichloroethane|1, 2 dichloroethane|1,2 -dichloroethane|1,2 dichloro ethane|1,2,-dichloroethane|1,2- dichloroethane|1,2-di-chloroethane|1,2-dichloro ethane|1.2-di-chloroethane|1,2-ethylenedichloride|ClCH2-CH2Cl|dichloro-1, 2-ethane|ethylene dichloride (1,2-dichloroethane)|C1CH2CH2Cl|ClCH2CH2C1|EDC, JMAF|Cl(CH2)2Cl|12-DICHLOROETHANE|Ethene, dichloro-(9CI)|bmse000568|1,2-Dichloroethane (OSHA)|CHEMBL16370|1,2-Dichloroethane ACS grade|1,2-Dichloroethane, for HPLC|Ethylene dichloride, BSI, ISO|34H - WFD H|1,2-Dichloroethane, ACS reagent|1,2-Dichloroethane, HPLC Grade|ETHYLENE DICHLORIDE [INCI]|Ethylene dichloride (ACGIH:OSHA)|Tox21_202466|1,2-Dichloroethane, LR, >=99%|NU-G00511|STL264187|AKOS000120021|1,2-Dichloroethane, p.a., 99.5%|DB03733|1,2-Dichloroethane, AR, >=99.5%|1,2-Dichloroethane, anhydrous, 99.8%|18A - Haloforms & C

hlorinated Solvents|NCGC00091763-01|NCGC00091763-02|NCGC00091763-03|NCGC00260015-01|06A - Haloforms and Chlorinated Solvents|CAS-107-06-2|1,2-Dichloroethane, for HPLC, 99.8%|1,2-Dichloroethane, ACS reagent, >=99%|1,2-Dichloroethane, ReagentPlus(R), 99%|1,2-Dichloroethane, for HPLC, >=99.8%|1,2-Dichloroethane, Spectrophotometric Grade|1ST000043-1000E|D0310|D0364|E0289|NS00004126|R 150|1,2-Dichloroethane 10 microg/mL in Methanol|EN300-19802|1,2-Dichloroethane, ACS reagent, >=99.0%|C06752|1,2-Dichloroethane 1000 microg/mL in Methanol|1,2-Dichloroethane, SAJ first grade, >=99.0%|Q161480|1,2-Dichloroethane, JIS special grade, >=99.5%|Ethylene dichloride [UN1184] [Flammable liquid]|J-503815|1,2-Dichloroethane, anhydrous, ZerO2(TM), 99.8%|1,2-Dichloroethane, spectrophotometric grade, >=99%|1,2-Dichloroethane Solution in Ethyl acetate, 1000ug/mL|1,2-Dichloroethane, puriss., absolute, over molecular sieve (H2O <=0.005%), >=99.5% (GC)",98.960,C2H4Cl2,0.000,6.000,1.500,4,0,0,1,InChI=1S/C2H4Cl2/c3-1-2-4/h1-2H2,C(CCl)Cl,C(CCl)Cl,WSLDOOZREJYCGB-UHFFFAOYSA-N,"1,2-dichloroethane",97.969,97.969,0,1,0,0,0,0,0,0,0,8712,46484,25937,"",Agrochemical Information|Biological Test Results|Interactions and Pathways|Chemical and Physical Properties|Classification|Drug and Medication Information|Food Additives and Ingredients|Identification|Literature|Patents|Pharmacology and Biochemistry|Safety and Hazards|Toxicity|Use and Manufacturing|Associated Disorders and Diseases|Spectral Information|Taxonomy|Biological Test Results: Active|Biological Test Results: Micromolar,19,421|426|427|433|434|435|445|530|540|541|542|543|544|545|546|584|585|595|596|603|605|654|655|656|657|658|659|660|661|662|663|664|665|666|667|875|880|881|884|885|886|887|889|892|893|894|900|902|912|921|923|924|925|926|938|946|947|948|955|960|961|962|963|964|965|966|967|968|969|970|971|972|973|974|975|976|977|978|979|980|981|982|983|984|985|986|987|988|989|993|994|995|1030|1188|1189|1194|1199|1205|1208|1452|1457|1458|1469|1471|1476|1477|1478|1479|1948|2101|2107|2112|2120|2517|2546|2549|2551|19262|19825|37562|101345|159270|162229|212400|485290|588209|588210|588513|588514|588515|588516|588526|588527|588532|588533|588534|588535|588536|588537|588541|588543|588544|588545|588546|588547|588834|603957|651631|651632|651633|651634|651741|651743|651749|651751|651754|651755|651757|651758|651777|651778|651802|651838|720516|720552|720634|720635|720636|720637|720652|720653|720659|720674|720675|720678|720679|720680|720681|720682|720683|720684|720685|720686|720687|720691|720692|720693|720719|720725|743012|743014|743015|743033|743035|743036|743040|743041|743042|743053|743054|743063|743064|743065|743066|743067|743069|743074|743075|743077|743078|743079|743080|743081|743083|743084|743085|743086|743091|743094|743122|743139|743140|743191|743194|743199|743202|743203|743209|743210|743211|743212|743213|743215|743217|743218|743219|743220|743221|743222|743223|743224|743225|743226|743227|743228|743239|743240|743241|743242|743288|743292|1159509|1159515|1159516|1159517|1159518|1159519|1159520|1159521|1159523|1159525|1159526|1159527|1159528|1159529|1159531|1159551|1159552|1159553|1159555|1224834|1224835|1224836|1224837|1224838|1224839|1224840|1224841|1224842|1224843|1224844|1224845|1224846|1224847|1224848|1224849|1224867|1224868|1224869|1224870|1224871|1224872|1224873|1224874|1224875|1224876|1224877|1224878|1224879|1224880|1224881|1224882|1224883|1224884|1224885|1224886|1224887|1224888|1224889|1224890|1224892|1224893|1224894|1224895|1224896|1259241|1259242|1259243|1259244|1259247|1259248|1259364|1259365|1259366|1259367|1259368|1259369|1259377|1259378|1259379|1259380|1259381|1259382|1259383|1259384|1259385|1259386|1259387|1259388|1259390|1259391|1259392|1259393|1259394|1259395|1259396|1259401|1259402|1259403|1259404|1259407|1259408|1259411|1346784|1346795|1346798|1346799|1346824|1346829|1346859|1346877|1346891|1346924|1346977|1346978|1346979|1346980|1346981|1346982|1347030|1347031|1347032|1347033|1347034|1347035|1347036|1347037|1347038|1347395|1347397|1347398|1347399|1671196|1671197|1671198|1671199|1671200|1671201|1745844|1919968|1919969|1919970|1920063|1920064|1920065|1920067|1920068|1963577|1963578|1963579|1963580|1963581|1963582|1963583|1963584|1963585,20040916,"001Chemical|10X CHEM|1st Scientific|3B Scientific (Wuhan) Corp|A2B Chem|AA BLOCKS|AAA Chemistry|abcr GmbH|ABI Chem|Acadechem|Achemica|Achemtek|Acmec Biochemical|AKos Consulting & Solutions|Amadis Chemical|Ambinter|AN PharmaTech|Angene Chemical|Anward|Apexmol|Aurora Fine Chemicals LLC|BenchChem|Bic Biotech|Biocore|BioCyc|Biological Magnetic Resonance Data Bank (BMRB)|Biosynth|BOC Sciences|Broad Institute|Cangzhou Enke Pharma Tech Co.,Ltd.|Carcinogenic Potency Database (CPDB)|ChEBI|Chem-Space.com Database|Chembase.cn|ChEMBL|ChemDB|Chemenu Inc.|ChemExper Chemical Directory|Chemical Carcinogenesis Research Infor

mation System (CCRIS)|ChemIDplus|Chemieliva Pharmaceutical Co., Ltd|ChemMol|ChemSpider|ChemTik|CHIRALEN|Clearsynth|Collaborative Drug Discovery, Inc.|Comparative Toxicogenomics Database (CTD)|Cooke Chemical Co., Ltd|Crystallography Open Database (COD)|CymitQuimica|DiscoveryGate|DrugBank|EAWAG Biocatalysis/Biodegradation Database|ECI Group, LCSB, University of Luxembourg|Egon Willighagen, Department of Bioinformatics - BiGCaT, Maastricht University|eMolecules|Enamine|enviPath|EPA DSSTox|EPA Substance Registry Services|EvitaChem|FDA Global Substance Registration System (GSRS)|Finetech Industry Limited|Fisher Chemical|Genetic Toxicology Data Bank (GENE-TOX)|Glentham Life Sciences Ltd.|Google Patents|Hairui Chemical|Human Metabolome Database (HMDB)|IBM|iChemical Technology USA Inc|InvivoChem|ISpharm|J&H Chemical Co.,ltd|Japan Chemical Substance Dictionary (Nikkaji)|KEGG|King Scientific|labseeker|Lan Pharmatech|LeadScope|LEAPCHEM|LGC Standards|MassBank Europe|MassBank of North America (MoNA)|Mcule|Metabolomics Workbench|MolCore|Molecule Market|Molepedia|MolGenie|MP Biomedicals|MuseChem|National Center for Advancing Translational Sciences (NCATS)|Nature Chemical Biology|NCBI Structure|NextBio|NextMove Software|NIST Chemistry WebBook|NIST Mass Spectrometry Data Center|NMRShiftDB|NORMAN Suspect List Exchange|NovoSeek|Oakwood Products|Parchem|PATENTSCOPE (WIPO)|Phion Ltd|Probes & Drugs portal|Protein Data Bank in Europe (PDBe)|PubChem Reference Collection|RR Scientific|Santa Cruz Biotechnology, Inc.|SCRIPDB|Sigma-Aldrich|Sinfoo Biotech|SLING Consortium|SMID|Smolecule|Springer Nature|SpringerMaterials|Starshine Chemical|SureChEMBL|SynQuest Laboratories|TCI (Tokyo Chemical Industry)|THE BioTek|Thermo Fisher Scientific|Thieme Chemistry|Thomson Pharma|Thoreauchem|TimTec|Tox21|ToxPlanet|Tractus|Vitas-M Laboratory|VladaChem|VWR, Part of Avantor|Wikidata|Wiley|Win-Win Chemical|Wubei-Biochem|Wutech|Yick-Vic Chemicals & Pharmaceuticals (HK) Ltd.|Yuhao Chemical|ZeroPM - Zero Pollution of Persistent, Mobile Substances|ZINC",Chemical Vendors|Curation Efforts|Governmental Organizations|Journal Publishers|Legacy Depositors|NIH Initiatives|Research and Development|Subscription Services,""

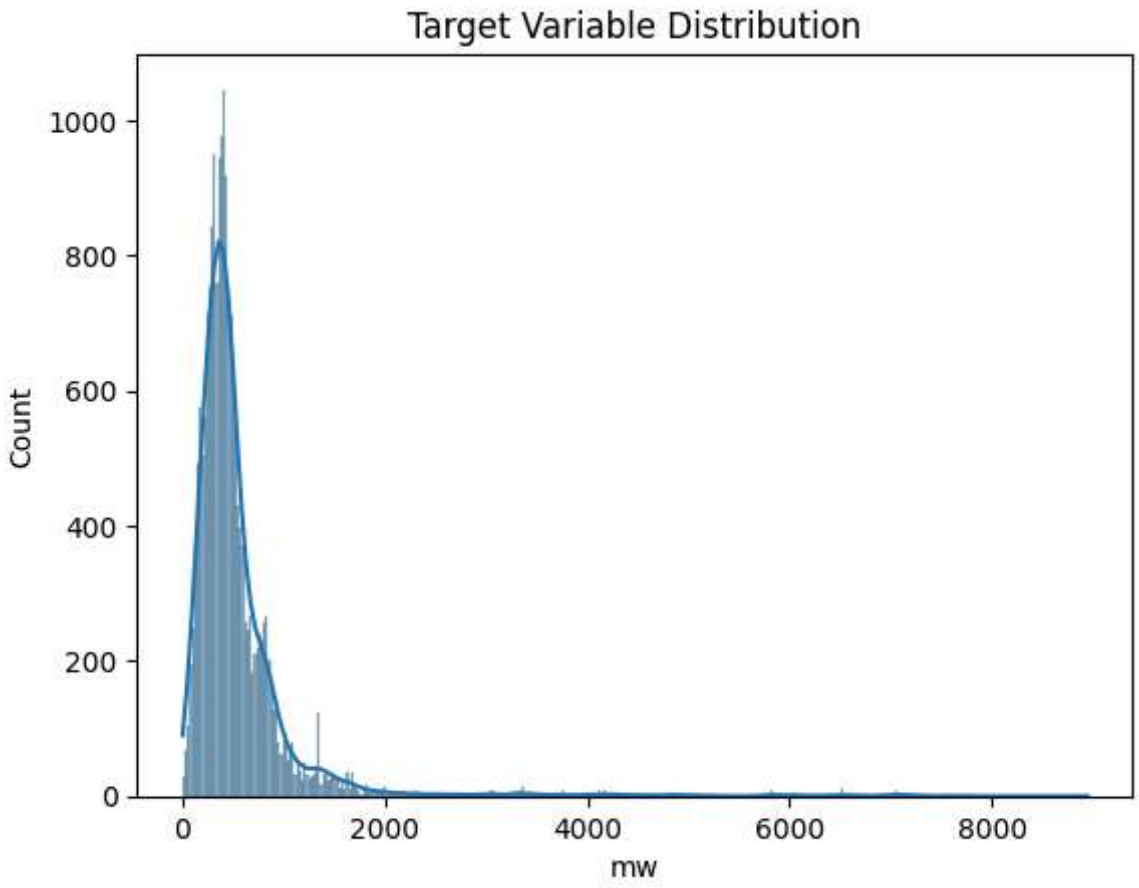
34,2-Chloroethanol,"2-chloroethanol|Ethylene chlorohydrin|107-07-3|Ethanol, 2-chloro-|Glycol chlorohydrin|2-Chloroethyl alcohol|Chloroethanol|Glycol monochlorohydrin|2-Monochloroethanol|Ethylchlorohydrin|2-Chlorethanol|2-Chloro-1-ethanol|2-Hydroxyethyl chloride|Glycomonochlorhydrin|beta-Chloroethyl alcohol|Ethylene chlorhydrin|Ethene, chlorohydrin|2-Chloorethanol|2-Chloraethanol|2-Cloroetanolo|beta-Chloroethanol|Aethylenechlorhydrin|2-Chloro-1-hydroxyethane|Glicol monocloridrina|Ethyleen-chloorhydrine|Glycolmonochloorhydrine|NCI-C50135|Ethylene glycol, chlorohydrin|beta-Hydroxyethyl chloride|Chloroalcohol|2-CHLORO-ETHANOL|Monochlorhydrine du glycol|.beta.-Chloroethyl alcohol|Ethanol, chloro-|NSC 122289|.beta.-Chloroethanol|MFCD00002829|.beta.-Hydroxyethyl chloride|753N66IHAN|59826-67-4|DTXSID1021877|CHEBI:28200|NSC-122289|delta-Chloroethanol|2-Chloorethanol [Dutch]|2-Chlorethanol [German]|2-Chloraethanol [German]|2-Cloroetanolo [Italian]|2-Chloro Ethanol|Aethylenechlorhydrin [German]|CCRIS 859|Ethyleen-chloorhydrine [Dutch]|Chloroethylowy alkohol|Glycolmonochloorhydrine [Dutch]|HSDB 426|Chloroethylowy alkohol [Polish]|Glicol monocloridrina [Italian]|Monochlorhydrine du glycol [French]|EINECS 203-459-7|UN1135|BRN 0878139|UNII-753N66IHAN|AI3-52326|2-chloranylethanol|ethylenechlorohydrin|delta-Chloroethanolchloroethylowy alkohol [Polish]|ethylenechlorohydrine|2-choroethan-1-ol|ethylene chlorohydrine|2-chloroethan-1-ol|2-Chloroethanol (ethylene chlorohydrin)|delta-Chloroethanolchloroethylowy alkohol|2-chloro-ethan-1-ol|CH2ClCH2OH|buta-1,3-diene-1,1,4-tricarboxylic acid|2-Chloroethanol, 99%|1-chloro-2-hydroxyethane|bmse000360|EC 203-459-7|WLN: Q2G|4-01-00-01372 (Beilstein Handbook Reference)|CHEMBL191244|DTXCID601877|2-CHLOROETHANOL [HSDB]|868 - Ethylene oxide in spices|ETHYLENE CHLOROHYDRIN [MI]|Tox21_300043|ArgoGel(TM)-Cl, 1 % cross-linked|BBL011477|NSC122289|STL146589|AKOS000119040|UN 1135|857 - Ethylene oxide in food products|NCGC00247890-01|NCGC00254022-01|CAS-107-07-3|.delta.-Chloroethanolchloroethylowy alkohol|2-Chloroethanol 100 microg/mL in Methanol|2-Chloroethanol 1000 microg/mL in Methanol|Ethylene chlorohydrin [UN1135] [Poison]|NS00003567|C06753|2-Chloroethanol, SAJ special grade, >=99.0%|InChI=1/C2H5ClO/c3-1-2-4/h4H,1-2H|A801572|Q209354|2-Chloroethanol, PESTANAL(R), analytical standard|2-Chloroethanol, puriss. p.a., >=99.0% (GC)|J-509022",80.510,C2H5ClO,20.200,10.000,-0.100,4,1,1,1,"InChI=1S/C2H5ClO/c3-1-2-4/h4H,1-2H2",C(CCl)O,C(CCl)O,SZIFAVKTNFCBPC-UHFFFAOY

SA-N,2-chloroethanol,80.0029,80.0029,0,1,0,0,0,0,0,0,0,1796,35304,17208,Ethylene Chlorohydrin,Biological Test Results|Interactions and Pathways|Chemical and Physical Properties|Classification|Drug and Medication Information|Identification|Literature|Patents|Pharmacology and Biochemistry|Safety and Hazards|Toxicity|Use and Manufacturing|Associated Disorders and Diseases|Spectral Information|Biological Test Results: Active,15,256|1188|384212|651631|651632|651633|651634|720516|720552|720634|720635|720637|720674|720675|720678|720679|720680|720681|720682|720683|720684|720685|720686|720687|720691|720692|720693|720719|720725|743012|743014|743015|743033|743035|743036|743040|743041|743042|743053|743054|743063|743064|743065|743066|743067|743069|743074|743075|743077|743078|743079|743080|743081|743083|743084|743085|743086|743091|743094|743122|743139|743140|743191|743194|743199|743202|743203|743209|743210|743211|743212|743213|743215|743217|743218|743219|743220|743221|743222|743223|743224|743225|743226|743227|743228|743239|743240|743241|743242|1094728|1159509|1159515|1159516|1159517|1159518|1159519|1159520|1159521|1159523|1159525|1159526|1159527|1159528|1159529|1159531|1159551|1159552|1159553|1159555|1224834|1224835|1224836|1224837|1224838|1224839|1224840|1224841|1224842|1224843|1224844|1224845|1224846|1224847|1224848|1224849|1224867|1224868|1224869|1224870|1224871|1224872|1224873|1224874|1224875|1224876|1224877|1224878|1224879|1224880|1224881|1224882|1224883|1224884|1224885|1224886|1224887|1224888|1224889|1224890|1224892|1224893|1224894|1224895|1224896|1259241|1259242|1259243|1259244|1259247|1259248|1259364|1259365|1259366|1259367|1259368|1259369|1259377|1259378|1259379|1259380|1259381|1259382|1259383|1259384|1259385|1259386|1259387|1259388|1259390|1259391|1259392|1259393|1259394|1259395|1259396|1259401|1259402|1259403|1259404|1259407|1259408|1259411|1346784|1346795|1346798|1346799|1346824|1346829|1346859|1346877|1346891|1346924|1346977|1346978|1346979|1346980|1346981|1346982|1347030|1347031|1347032|1347033|1347034|1347035|1347036|1347037|1347038|1347395|1347397|1347398|1347399|1671196|1671197|1671198|1671199|1671200|1671201|1919968|1919969|1919970|1920063|1920064|1920065|1920067|1920068|1963577|1963578|1963579|1963580|1963581|1963582|1963583|1963584|1963585,20050326,"10X CHEM|3B Scientific (Wuhan) Corp|AA BLOCKS|abcr GmbH|ABI Chem|Acadechem|Achemica|Achemtek|ACT Chemical|AK Scientific, Inc. (AKSCI)|AKos Consulting & Solutions|Alfa Chemistry|Amadis Chemical|Ambinter|AN PharmaTech|Angene Chemical|AZEPINE|BenchChem|Bic Biotech|BioCyc|Biological Magnetic Resonance Data Bank (BMRB)|Biosynth|BOC Sciences|ChEBI|Chem-Space.com Database|Chembase.cn|ChEMBL|ChemDB|ChemExper Chemical Directory|Chemical Carcinogenesis Research Information System (CCRIS)|ChemIDplus|Chemieliva Pharmaceutical Co., Ltd|ChemMol|ChemSpider|ChemTik|Comparative Toxicogenomics Database (CTD)|CymitQuimica|Discovery Gate|DTP/NCI|EAWAG Biocatalysis/Biodegradation Database|Egon Willighagen, Department of Bioinformatics - BiGCaT, Maastricht University|enviPath|EPA DSSTox|EPA Substance Registry Services|EvitaChem|FDA Global Substance Registration System (GSRS)|Genetic Toxicology Data Bank (GENE-TOX)|Glentham Life Sciences Ltd.|Google Patents|Human Metabolome Database (HMDB)|IBM|iChemical Technology USA Inc|ISpharm|IUPAC Digitized pKa Dataset|J&H Chemical Co.,ltd|Japan Chemical Substance Dictionary (Nikkaji)|KEGG|LeadScope|LGC Standards|Mcule|Metabolomics Workbench|Molepedia|MolGenie|MolPort|MuseChem|National Center for Advancing Translational Sciences (NCATS)|NextBio|NextMove Software|NIST Chemistry WebBook|NIST Mass Spectrometry Data Center|NMRShiftDB|NORMAN Suspect List Exchange|NovoSeek|Oakwood Products|OtavaChemicals|PATENTSCOPE (WIPO)|PubChem Reference Collection|RR Scientific|SCRIPDB|Sigma-Aldrich|Sinfoo Biotech|SLING Consortium|Smolecule|Springer Nature|SpringerMaterials|Starshine Chemical|SureChEMBL|THE BioTek|Thermo Fisher Scientific|Thieme Chemistry|Thomson Pharma|Tox21|ToxPlanet|Vitas-M Laboratory|VladaChem|Wikidata|Wiley|Wubei-Biochem|Wutech|Yick-Vic Chemicals & Pharmaceuticals (HK) Ltd.|Yuhao Chemical|ZeroPM - Zero Pollution of Persistent, Mobile Substances|ZINC",Chemical Vendors|Curation Efforts|Governmental Organizations|Journal Publishers|Legacy Depositors|NIH Initiatives|Research and Development|Subscription Services,""

Data Cleaning and Preparation

In [6]:
```
import pandas as pd
from sklearn.model_selection import train_test_split
```

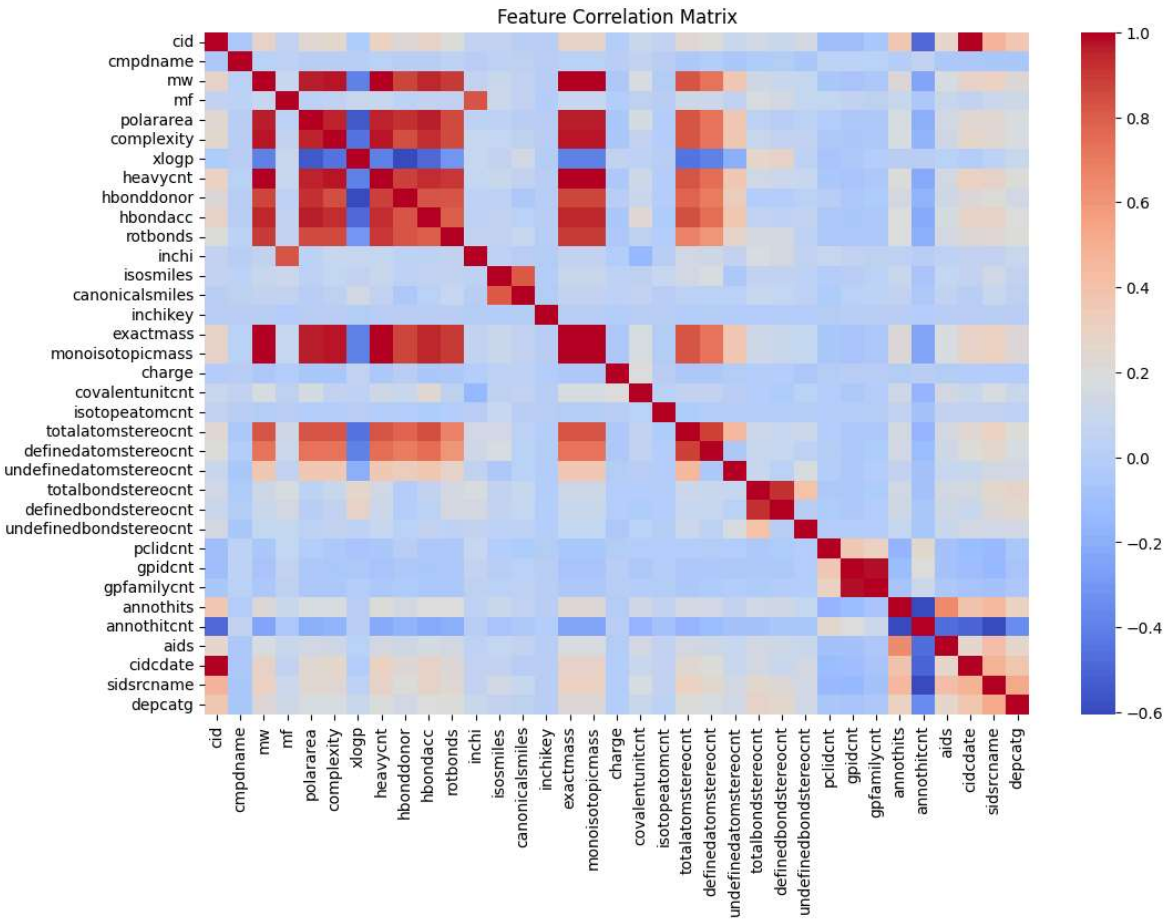
```python
from sklearn.preprocessing import StandardScaler, LabelEncoder
import seaborn as sns
import matplotlib.pyplot as plt

# Load dataset
file_path = r"C:\Users\USUARIO\Desktop\projectAD\PubChem_compound_cache_p-AA8AaJ
data = pd.read_csv(file_path)

# Step 1: Handle Missing Values
data.fillna(data.median(numeric_only=True), inplace=True)  # Fill numeric column
data.fillna("Unknown", inplace=True)  # Fill non-numeric columns with 'Unknown'

# Step 2: Drop Irrelevant Columns
columns_to_drop = ['cmpdsynonym', 'iupacname', 'meshheadings', 'annotation']
data.drop(columns=columns_to_drop, inplace=True, errors='ignore')

# Step 3: Label Encode Categorical Variables
categorical_columns = data.select_dtypes(include=['object']).columns
for col in categorical_columns:
    data[col] = LabelEncoder().fit_transform(data[col])

# Step 4: Select Features and Target
# Replace 'annotation' with your actual target column if applicable
target_column = 'mw'  # Example: use molecular weight for testing (update with y
if target_column not in data.columns:
    raise ValueError("Specify a valid target column.")
X = data.drop(columns=[target_column])
y = data[target_column]

# Step 5: Train-Test Split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_

# Step 6: Feature Scaling
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Step 7: Visualize the Target Variable
sns.histplot(y, kde=True)
plt.title("Target Variable Distribution")
plt.xlabel(target_column)
plt.show()
```



Identify Relevant Features

In [7]:

```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.feature_selection import SelectKBest, f_regression

# Correlation Analysis
correlation_matrix = data.corr()
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=False, cmap='coolwarm')
plt.title("Feature Correlation Matrix")
plt.show()

# Correlation with the Target Variable
target_corr = correlation_matrix['mw'].sort_values(ascending=False)
print("Correlation with Target (mw):\n", target_corr)

# Feature Selection with ANOVA F-test
X_selected = SelectKBest(score_func=f_regression, k=10).fit(X, y)
selected_features_indices = X_selected.get_support(indices=True)
selected_features = X.columns[selected_features_indices]
print("\nTop 10 Features Selected:\n", selected_features)
```


Feature Correlation Matrix
cid
cmpdname
mw
mf
polararea
complexity
xlogp
heavycnt
hbonddonor
hbondacc
rotbonds
inchi
isosmiles
canonicalsmiles
inchikey
exactmass
monoisotopicmass
charge
covalentunitcnt
isotopeatomcnt
totalatomstereocnt
definedatomstereocnt
undefinedatomstereocnt
totalbondstereocnt
definedbondstereocnt
undefinedbondstereocnt
pclidcnt
gpidcnt
gpfamilycnt
annothits
annothitcnt
aids
cidcdate
sidsrcname
depcatg
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

```
Correlation with Target (mw):
 mw                          1.000000
exactmass                   1.000000
monoisotopicmass            1.000000
heavycnt                    0.993832
complexity                  0.969211
polararea                   0.958148
hbondacc                    0.942022
rotbonds                    0.891398
hbonddonor                  0.867478
totalatomstereocnt          0.827521
definedatomstereocnt        0.735767
undefinedatomstereocnt      0.355534
sidsrcname                  0.293213
cidcdate                    0.291723
cid                         0.284835
annothits                   0.213014
depcatg                     0.210573
covalentunitcnt             0.167048
aids                        0.161339
totalbondstereocnt          0.122347
definedbondstereocnt        0.099505
isosmiles                   0.088806
mf                          0.082893
undefinedbondstereocnt      0.077354
inchi                       0.064914
canonicalsmiles             0.061516
cmpdname                    0.021213
inchikey                    0.003780
isotopeatomcnt             -0.005714
charge                     -0.042492
gpfamilycnt                -0.045346
pclidcnt                   -0.049972
gpidcnt                    -0.066208
annothitcnt                -0.229280
xlogp                      -0.400436
Name: mw, dtype: float64

Top 10 Features Selected:
 Index(['polararea', 'complexity', 'heavycnt', 'hbonddonor', 'hbondacc',
       'rotbonds', 'exactmass', 'monoisotopicmass', 'totalatomstereocnt',
       'definedatomstereocnt'],
      dtype='object')
```

Top Relevant Features (Based on Correlation Matrix)

polararea: Likely to have a strong correlation with mw.

complexity: Molecular complexity may relate strongly to molecular weight.

heavycnt: Number of heavy atoms is directly related to molecular weight.

rotbonds: More rotatable bonds might indicate larger molecules.

exactmass: This will naturally correlate with mw as they measure similar properties.

Feature Selection

Select Features Based on High Correlation:

In [8]:
```
# Correlation with target variable (mw)
threshold = 0.5  # Adjust threshold as needed
high_corr_features = target_corr[target_corr.abs() > threshold].index
print("Features highly correlated with 'mw':\n", high_corr_features)
```

```
Features highly correlated with 'mw':
 Index(['mw', 'exactmass', 'monoisotopicmass', 'heavycnt', 'complexity',
       'polararea', 'hbondacc', 'rotbonds', 'hbonddonor', 'totalatomstereocnt',
       'definedatomstereocnt'],
      dtype='object')
```

Model Building

In [9]:
```
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

# Select the highly correlated features
selected_features = [
    'exactmass', 'monoisotopicmass', 'heavycnt', 'complexity', 'polararea',
    'hbondacc', 'rotbonds', 'hbonddonor', 'totalatomstereocnt', 'definedatomster
]

X = data[selected_features]
y = data['mw']

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_

# Train a Random Forest Regressor
model = RandomForestRegressor(random_state=42)
model.fit(X_train, y_train)

# Predictions
y_pred = model.predict(X_test)

# Evaluate the model
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"Model Evaluation:\nMAE: {mae}\nMSE: {mse}\nR²: {r2}")
```

```
Model Evaluation:
MAE: 0.46170927939325873
MSE: 6.041973965414473
R²: 0.9999850721503823
```

Model Performance

The results indicate an excellent model fit:

MAE (Mean Absolute Error): ~0.46

The average error between the predicted and actual molecular weights is very low.

MSE (Mean Squared Error): ~6.04

This low value indicates very small deviations on average, with larger errors penalized more heavily.

R^2 (Coefficient of Determination): ~0.99999

Almost perfect explanation of the variance in the target variable (mw) by the selected features.

Analysis

The model is performing exceptionally well, likely due to strong correlations between mw and the selected features (exactmass, monoisotopicmass, heavycnt, etc.).

This suggests that the feature set is highly predictive for molecular weight.

Feature Importance and Cross-Validation

To further evaluate the model and gain insights into feature contributions, let's:

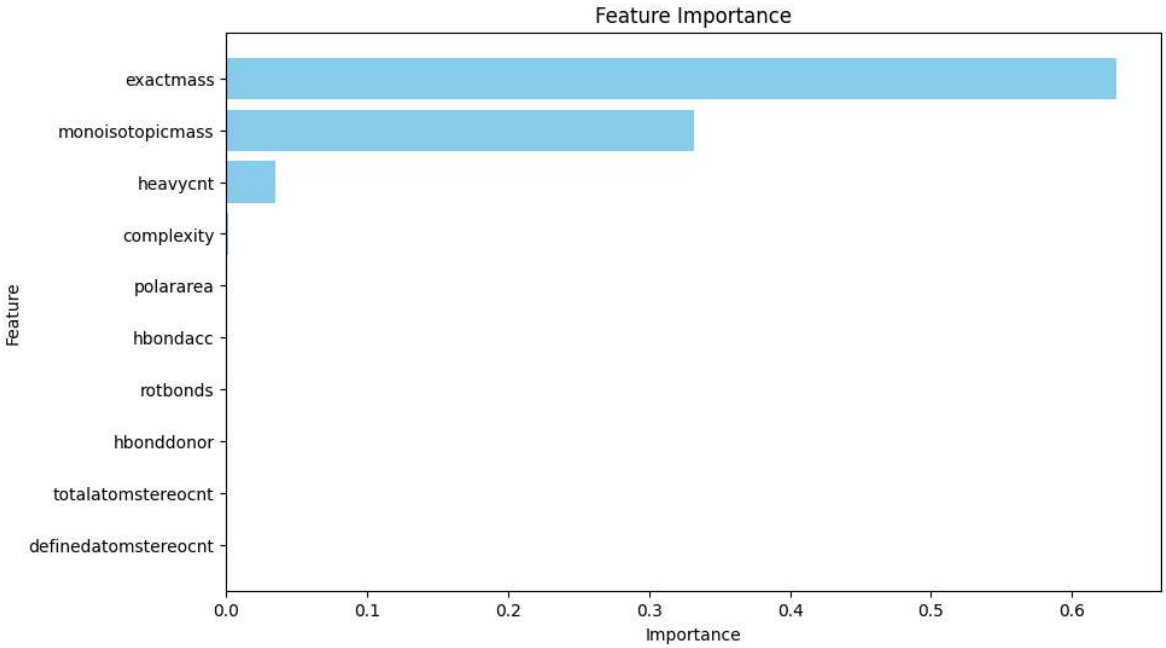
Visualize Feature Importance:

Determine which features contribute the most to the model's predictions.

Cross-Validation:

Perform cross-validation to ensure the model is robust and generalizes well.

In [10]:
```
from sklearn.model_selection import cross_val_score
import numpy as np
import matplotlib.pyplot as plt

# Feature Importance (for Random Forest Regressor)
importances = model.feature_importances_
feature_names = selected_features

# Plot feature importance
plt.figure(figsize=(10, 6))
plt.barh(feature_names, importances, color="skyblue")
plt.title("Feature Importance")
plt.xlabel("Importance")
plt.ylabel("Feature")
plt.gca().invert_yaxis()
plt.show()

# Cross-Validation
cv_scores = cross_val_score(model, X, y, scoring='r2', cv=5)  # 5-fold cross-val
print("Cross-Validation R² Scores:", cv_scores)
print("Mean R² Score:", np.mean(cv_scores))
```




```
Cross-Validation R² Scores: [0.99999317 0.99999556 0.99998398 0.99984888 0.999515
71]
Mean R² Score: 0.999867460229068
```

Analysis of Feature Importance

From the plot:

exactmass and monoisotopicmass are the most influential features, contributing significantly to the model's predictions. This is expected as they are directly related to

molecular weight.

heavycnt (number of heavy atoms) also has a meaningful impact, which makes sense chemically.

Other features like complexity, polararea, rotbonds, etc., have minimal impact.

This suggests that most of the predictive power comes from a small subset of features.

Cross-Validation Results

The cross-validation results confirm the model's robustness and excellent generalization capability:

R^2 Scores Across Folds:

Scores are consistently very high, ranging from 0.99951571 to 0.99999556.

This shows the model explains almost all the variance in the molecular weight (mw) across different data splits.

Mean R^2 Score:

0.999867460229068, confirming near-perfect performance on average across all folds.

Insights

Model Performance:

The Random Forest model is highly effective for predicting molecular weight using the selected features.

The consistent cross-validation scores indicate that the model is not overfitting and can generalize well to unseen data.

Feature Importance:

The dominant importance of exactmass and monoisotopicmass aligns with their chemical relevance to molecular weight.

Features like heavycnt further explain variations, suggesting that molecular weight is primarily influenced by atomic composition.

Actionable Insights

Focus on Exact Mass and Monoisotopic Mass:

For accurate molecular weight predictions, these two features should be prioritized.

Organizations can leverage these features for efficient molecular screening and optimization.

Heavy Atom Contribution:

The heavycnt feature (number of heavy atoms) provides a secondary layer of influence. This insight can guide compound design to target specific weight ranges.

Streamlined Feature Set:

Only a handful of features (exactmass, monoisotopicmass, heavycnt) are necessary for accurate predictions, enabling faster and more efficient model deployment.

Saving the Model and Predictions

In [11]:
```
import joblib
import pandas as pd

# Save the trained model
model_filename = "random_forest_mw_model.pkl"
joblib.dump(model, model_filename)
print(f"Model saved as {model_filename}")

# Make predictions and save to CSV
predictions = pd.DataFrame({
    "Actual": y_test,
    "Predicted": y_pred
})
predictions_filename = "mw_predictions.csv"
predictions.to_csv(predictions_filename, index=False)
print(f"Predictions saved as {predictions_filename}")
```

```
Model saved as random_forest_mw_model.pkl
Predictions saved as mw_predictions.csv
```

Load the Saved Model and Predict on New Data

If you want to proceed with making predictions on new data:

In [12]:
```python
# Load the saved model
loaded_model = joblib.load("random_forest_mw_model.pkl")
print("Model successfully loaded!")

# Example new data (replace with your actual dataset)
new_data = pd.DataFrame({
    "exactmass": [203.116, 97.969, 80.0029],
    "monoisotopicmass": [203.116, 97.969, 80.0029],
    "heavycnt": [14, 4, 4],
    "complexity": [214.0, 6.0, 10.0],
    "polararea": [66.4, 0.0, 20.2],
    "hbondacc": [4, 0, 1],
    "rotbonds": [5, 1, 1],
    "hbonddonor": [0, 0, 1],
    "totalatomstereocnt": [1, 0, 0],
    "definedatomstereocnt": [1, 0, 0]
})

# Predict molecular weights
new_predictions = loaded_model.predict(new_data)

# Add predictions to the DataFrame
new_data["Predicted_Molecular_Weight"] = new_predictions
print(new_data)

# Save predictions to a CSV file
new_predictions_filename = "new_mw_predictions.csv"
new_data.to_csv(new_predictions_filename, index=False)
print(f"New predictions saved as {new_predictions_filename}")
```

```
Model successfully loaded!
   exactmass  monoisotopicmass  heavycnt  complexity  polararea  hbondacc  \
0   203.1160          203.1160        14       214.0       66.4         4
1    97.9690           97.9690         4         6.0        0.0         0
2    80.0029           80.0029         4        10.0       20.2         1

   rotbonds  hbonddonor  totalatomstereocnt  definedatomstereocnt  \
0         5           0                   1                     1
1         1           0                   0                     0
2         1           1                   0                     0

   Predicted_Molecular_Weight
0                   203.35380
1                    98.82592
2                    80.51666
New predictions saved as new_mw_predictions.csv
```

Analyzing Predictions and Insights

Since the model has made predictions on new data and saved them in new_mw_predictions.csv, the next logical steps include:

Error Analysis (if actual molecular weights are available):

Compare predicted weights to actual weights (if provided) to assess model accuracy.

Calculate metrics like Mean Absolute Error (MAE) and Mean Percentage Error (MPE).

Extract Insights from Predictions and Feature Importance:

Analyze the relationship between key features (like exactmass, heavycnt, etc.) and the predicted molecular weights.

Visualize trends using scatterplots or feature distributions.

Error Analysis and Visualization

If you have actual weights for comparison:

In [13]:
```
# Assuming `actual_mw` column exists in the new data for comparison
new_data['Actual_Molecular_Weight'] = [203.116, 97.969, 80.0029]  # Replace with

# Calculate error metrics
new_data['Absolute_Error'] = abs(new_data['Actual_Molecular_Weight'] - new_data[
new_data['Percentage_Error'] = (new_data['Absolute_Error'] / new_data['Actual_Mo

mae = new_data['Absolute_Error'].mean()
mpe = new_data['Percentage_Error'].mean()

print(f"Mean Absolute Error (MAE): {mae}")
print(f"Mean Percentage Error (MPE): {mpe}%")

# Save analysis results
error_analysis_filename = "error_analysis.csv"
new_data.to_csv(error_analysis_filename, index=False)
print(f"Error analysis saved as {error_analysis_filename}")
```

```
Mean Absolute Error (MAE): 0.5361600000000095
Mean Percentage Error (MPE): 0.5446458423263804%
Error analysis saved as error_analysis.csv
```

Insights and Trends

In [14]:
```
import matplotlib.pyplot as plt
import seaborn as sns

# Scatter plot of exact mass vs predicted molecular weight
plt.figure(figsize=(8, 6))
sns.scatterplot(x='exactmass', y='Predicted_Molecular_Weight', data=new_data)
plt.title("Exact Mass vs Predicted Molecular Weight")
plt.xlabel("Exact Mass")
plt.ylabel("Predicted Molecular Weight")
plt.show()

# Relationship between heavycnt and predicted molecular weight
plt.figure(figsize=(8, 6))
sns.barplot(x='heavycnt', y='Predicted_Molecular_Weight', data=new_data)
plt.title("Heavy Atom Count vs Predicted Molecular Weight")
plt.xlabel("Heavy Atom Count")
```

```
plt.ylabel("Predicted Molecular Weight")
plt.show()
```