# درباره دیتاست

دیتاست شامل ۳۲۷۶ سطر و ۱۰ ستون است که در شکل زیر بعضی از ویژگی های آماری دیتاست نشان داده شده است.

| | ph | Hardness | Solids | Chloramines | Sulfate | Conductivity | Organic_carbon | Trihalomethanes | Turbidity | Potability |
|---|---|---|---|---|---|---|---|---|---|---|
| count | 2785.000000 | 3276.000000 | 3276.000000 | 3276.000000 | 2495.000000 | 3276.000000 | 3276.000000 | 3114.000000 | 3276.000000 | 3276.000000 |
| mean | 7.080795 | 196.369496 | 22014.092526 | 7.122277 | 333.775777 | 426.205111 | 14.284970 | 66.396293 | 3.966786 | 0.390110 |
| std | 1.594320 | 32.879761 | 8768.570828 | 1.583085 | 41.416840 | 80.824064 | 3.308162 | 16.175008 | 0.780382 | 0.487849 |
| min | 0.000000 | 47.432000 | 320.942611 | 0.352000 | 129.000000 | 181.483754 | 2.200000 | 0.738000 | 1.450000 | 0.000000 |
| 25% | 6.093092 | 176.850538 | 15666.690297 | 6.127421 | 307.699498 | 365.734414 | 12.065801 | 55.844536 | 3.439711 | 0.000000 |
| 50% | 7.036752 | 196.967627 | 20927.833607 | 7.130299 | 333.073546 | 421.884968 | 14.218338 | 66.622485 | 3.955028 | 0.000000 |
| 75% | 8.062066 | 216.667456 | 27332.762127 | 8.114887 | 359.950170 | 481.792304 | 16.557652 | 77.337473 | 4.500320 | 1.000000 |
| max | 14.000000 | 323.124000 | 61227.196008 | 13.127000 | 481.030642 | 753.342620 | 28.300000 | 124.000000 | 6.739000 | 1.000000 |

که به ترتیب هر ستون نشان دهنده عوامل تاثیرگذار بر کیفیت آب استخر ( Potability ) می باشند:

ph : ph آب ( ۰ تا ۱۴ )

Hardness : ظرفیت آب برای رسوب صابون بر حسب میلی گرم در لیتر

Solids : کل مواد جامد محلول در ppm

Chloramines : مقدار کلرامین در ppm

Solfate : مقدار سولفات های محلول بر حسب میلی گرم در لیتر

Conductivity : هدایت الکتریکی آب بر حسب µS/cm

Organic_carbon : مقدار کربن آلی در ppm

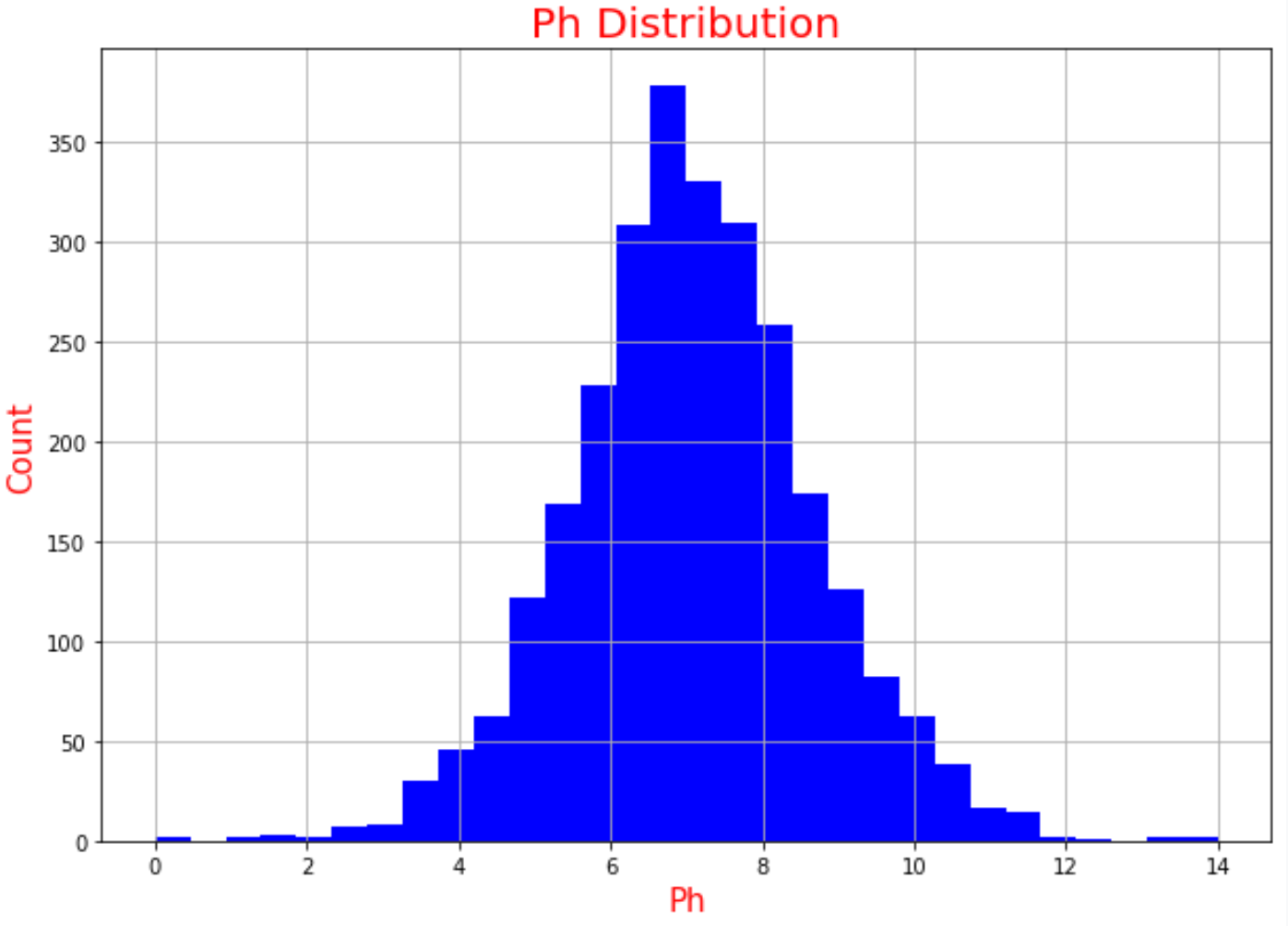
Trihalomethanes : مقدار تری هالومتان ها بر حسب میکروگرم در لیتر

Turbidity : اندازه گیری خاصیت ساطع نور آب در NTU

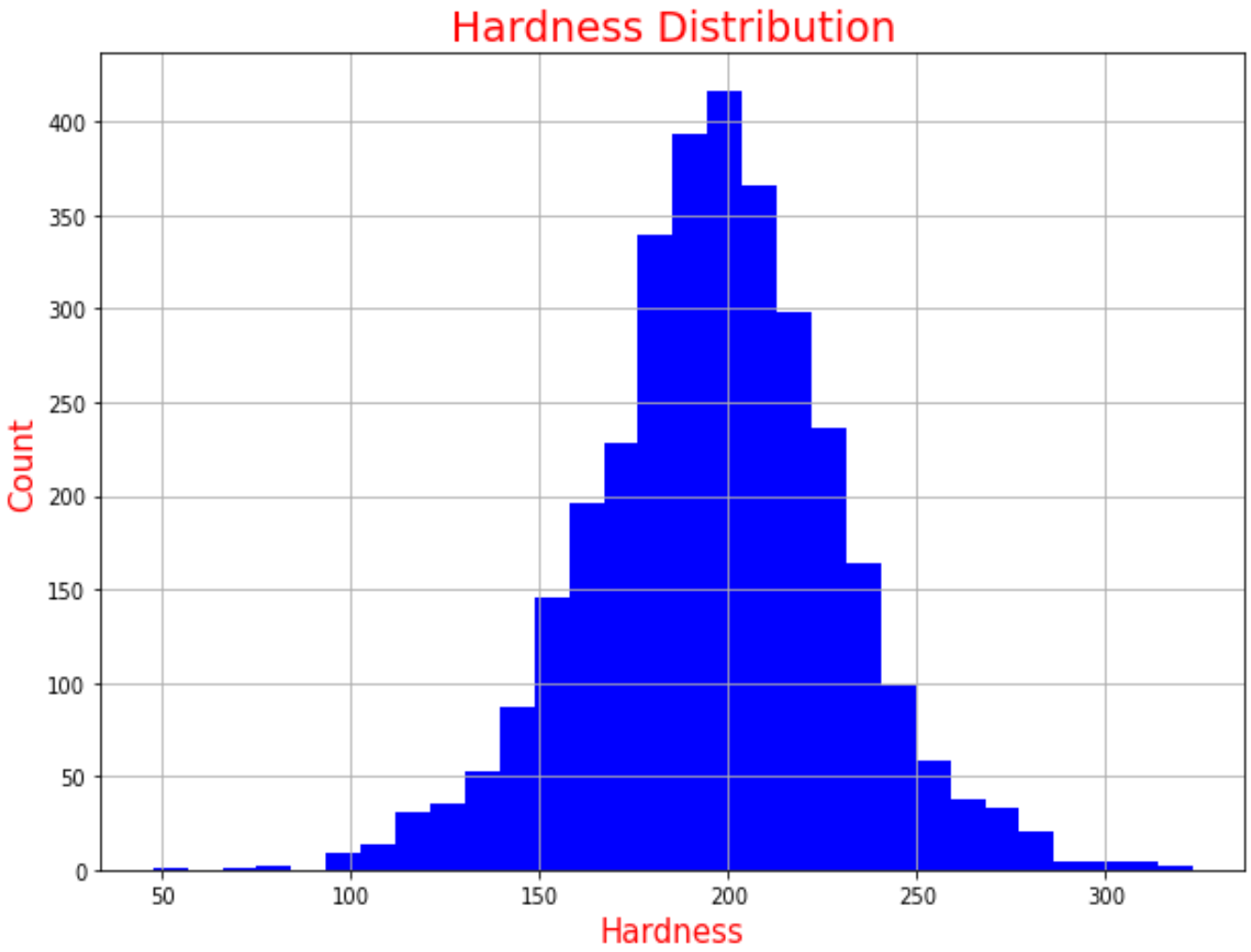
Potability : نشان می دهد که آیا آب برای مصرف انسان ایمن است یا خیر. قابل مصرف (۱) و غیر قابل مصرف (۰)

## مرحله اول : پیش پردازش دیتاست

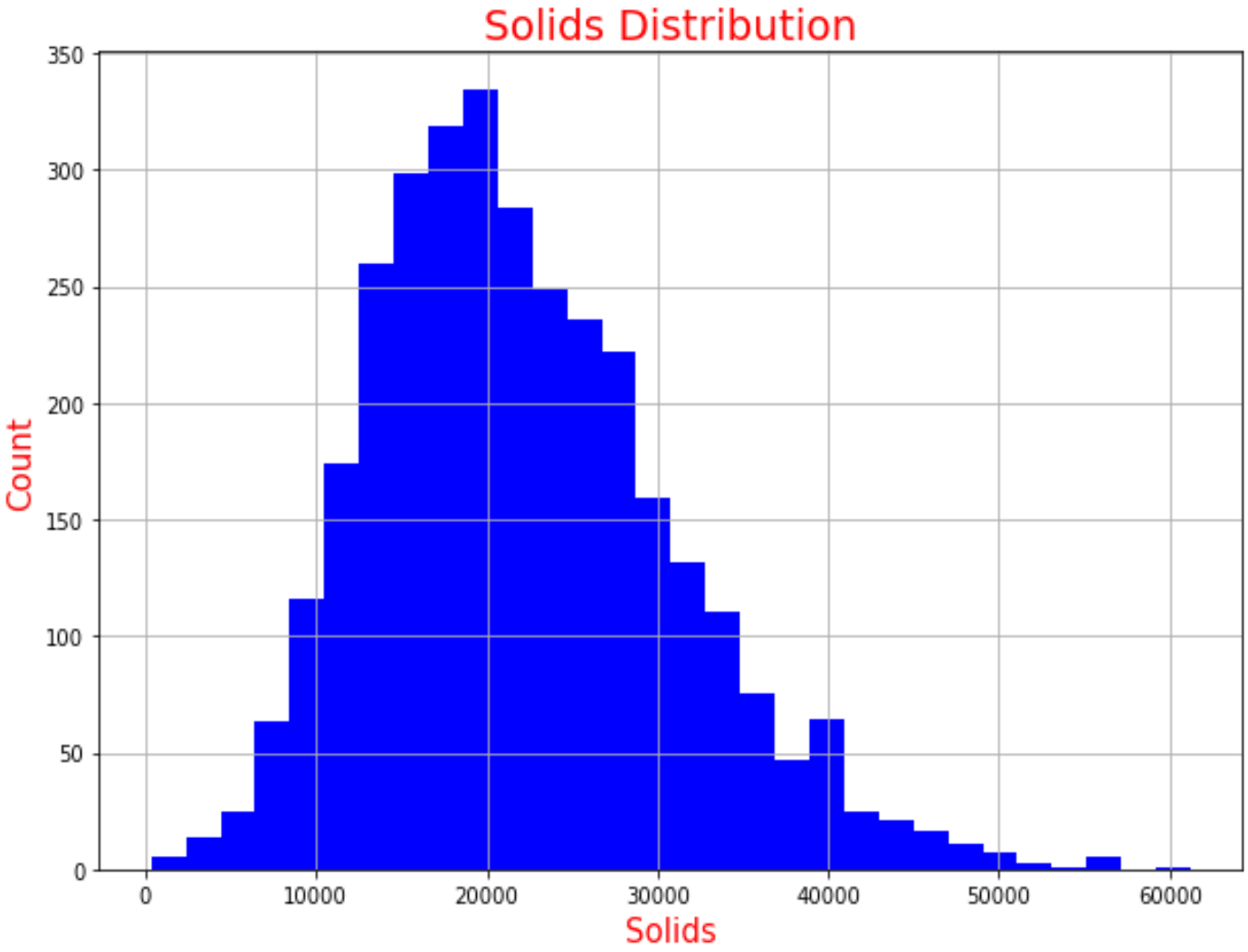
بررسی توزیع دیتاها در ستون ها از طریق نمودار هیستوگرام :



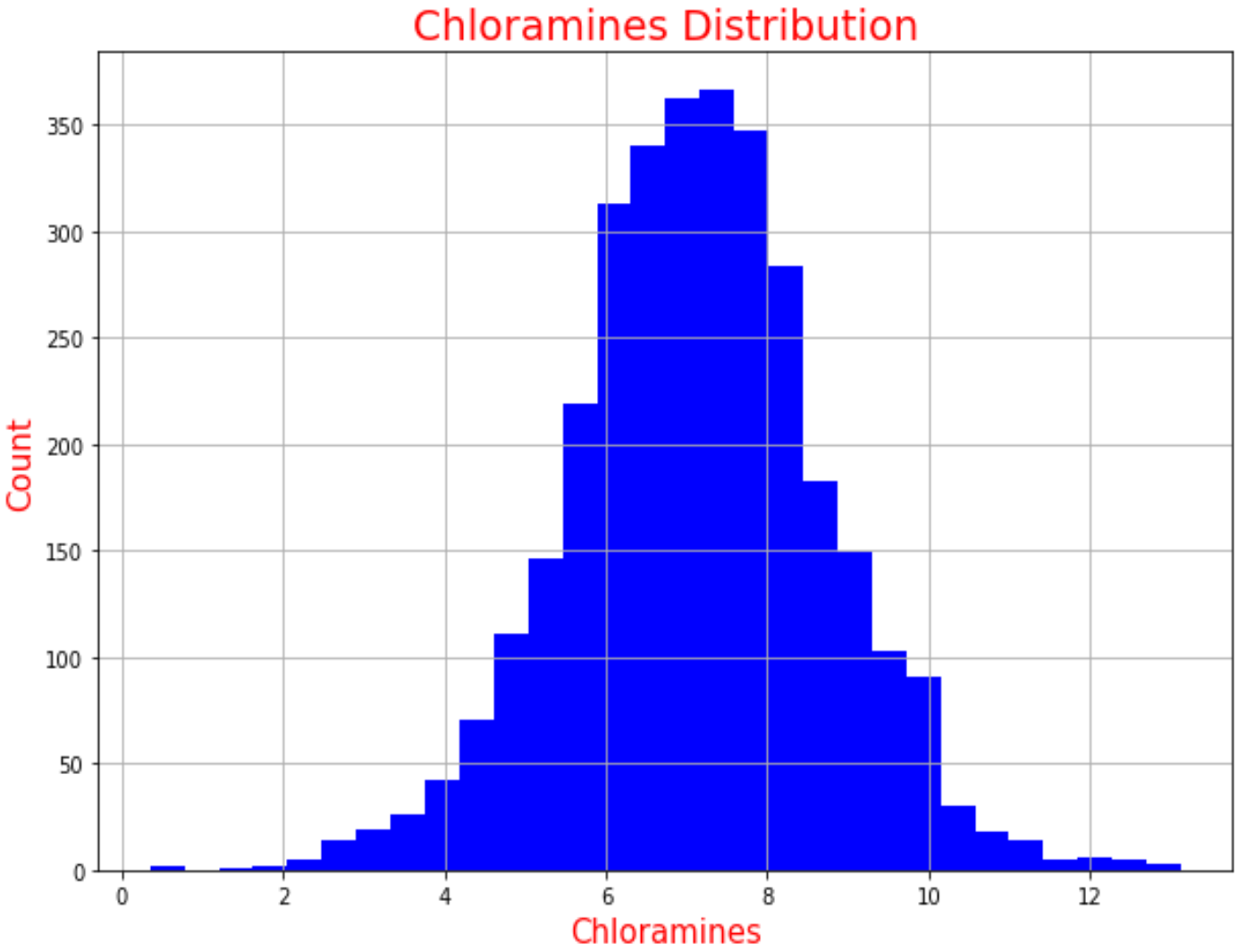

این نمودار نشان دهنده توزیع نرمال ph آب است. همانطور که مشخص است اغلب دیتاها در بازه ۶ تا ۸ که مناسب ترین بازه برای مصرف شناگران است ثبت شده است همچنین هرچه از مقدار متوسط ۷ فاصله گرفته شود آب غیرقابل مصرف و چگالی دیتاها کمتر میشود.



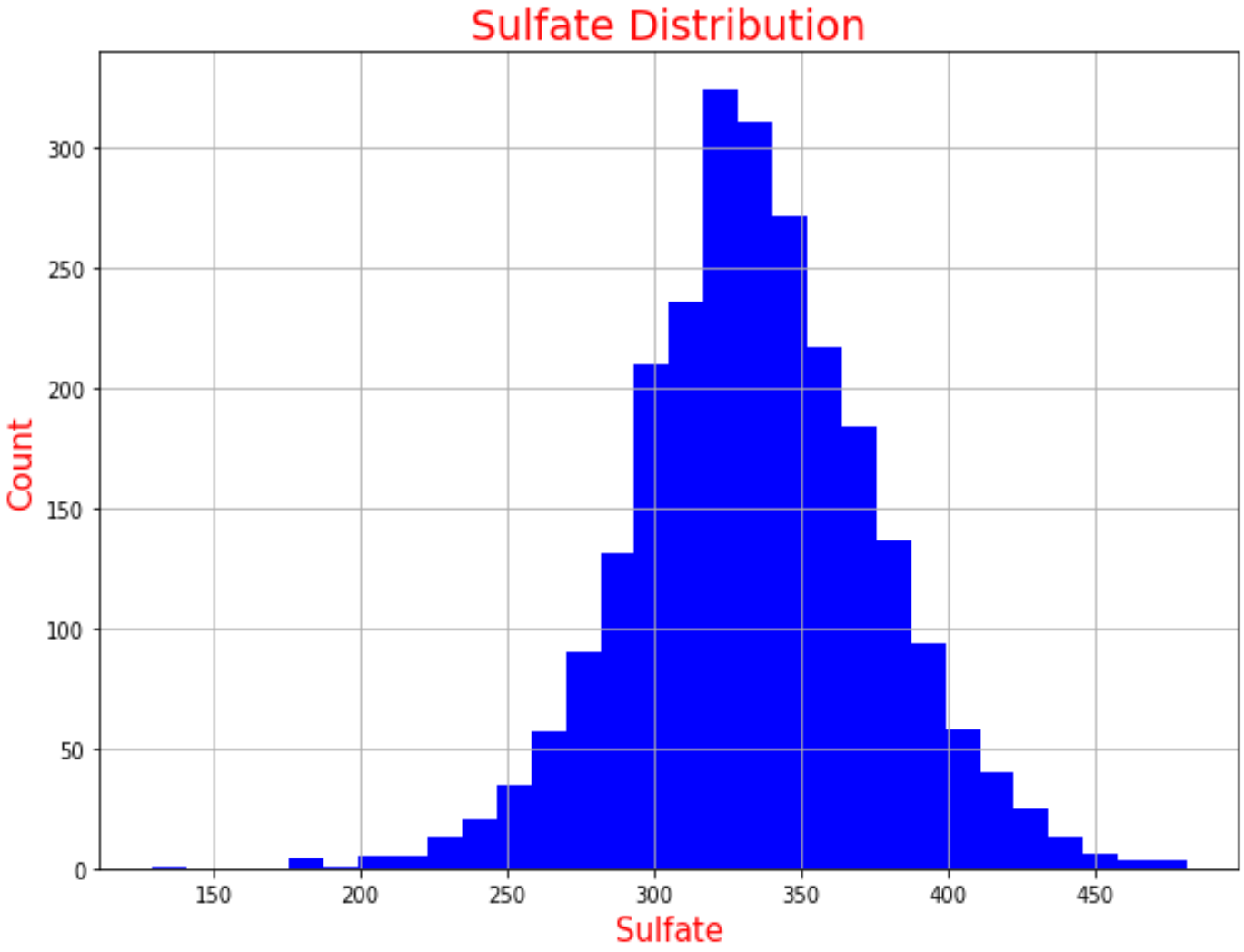

این نمودار نشان دهنده میزان سختی آب است. همانطور که مشخص است اغلب دیتاها در بازه ۱۸۰ تا ۲۲۰ ثبت شده است که نشان دهنده میزان متوسط سختی آب برای استخر است همچنین بیشترین دیتاهای ثبت شده در میزان سختی ۲۰۰ است که مناسب ترین میزان برای آب استخر است.



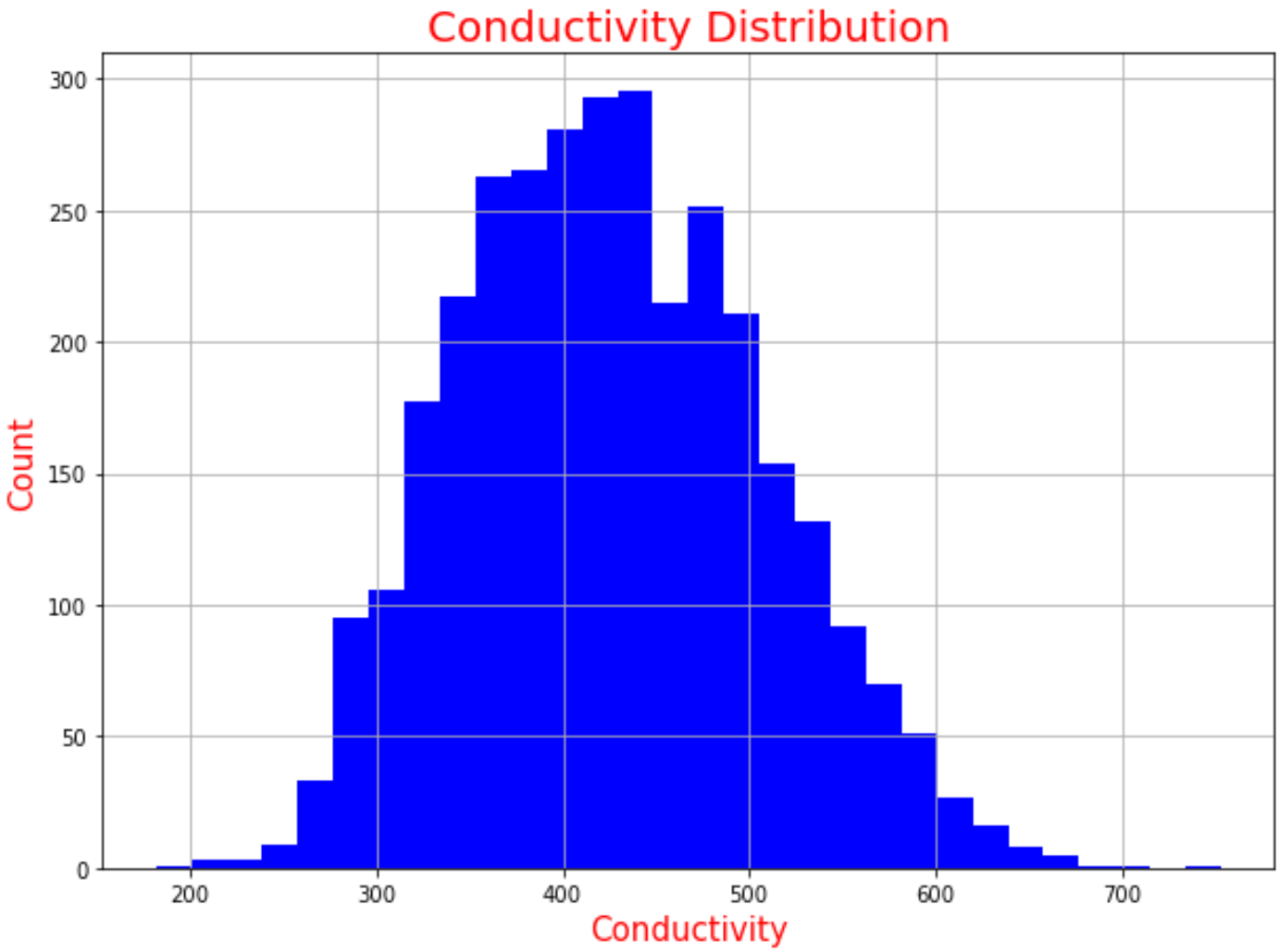

این نمودار نشان دهنده میزان کل مواد جامد در آب بر حسب ppm است که بیشترین تعداد دیتا ها در مقدار ۲۰هزار ثبت شده است و بیشترین میزان آن در ۶۰ هزار ثبت شده است.



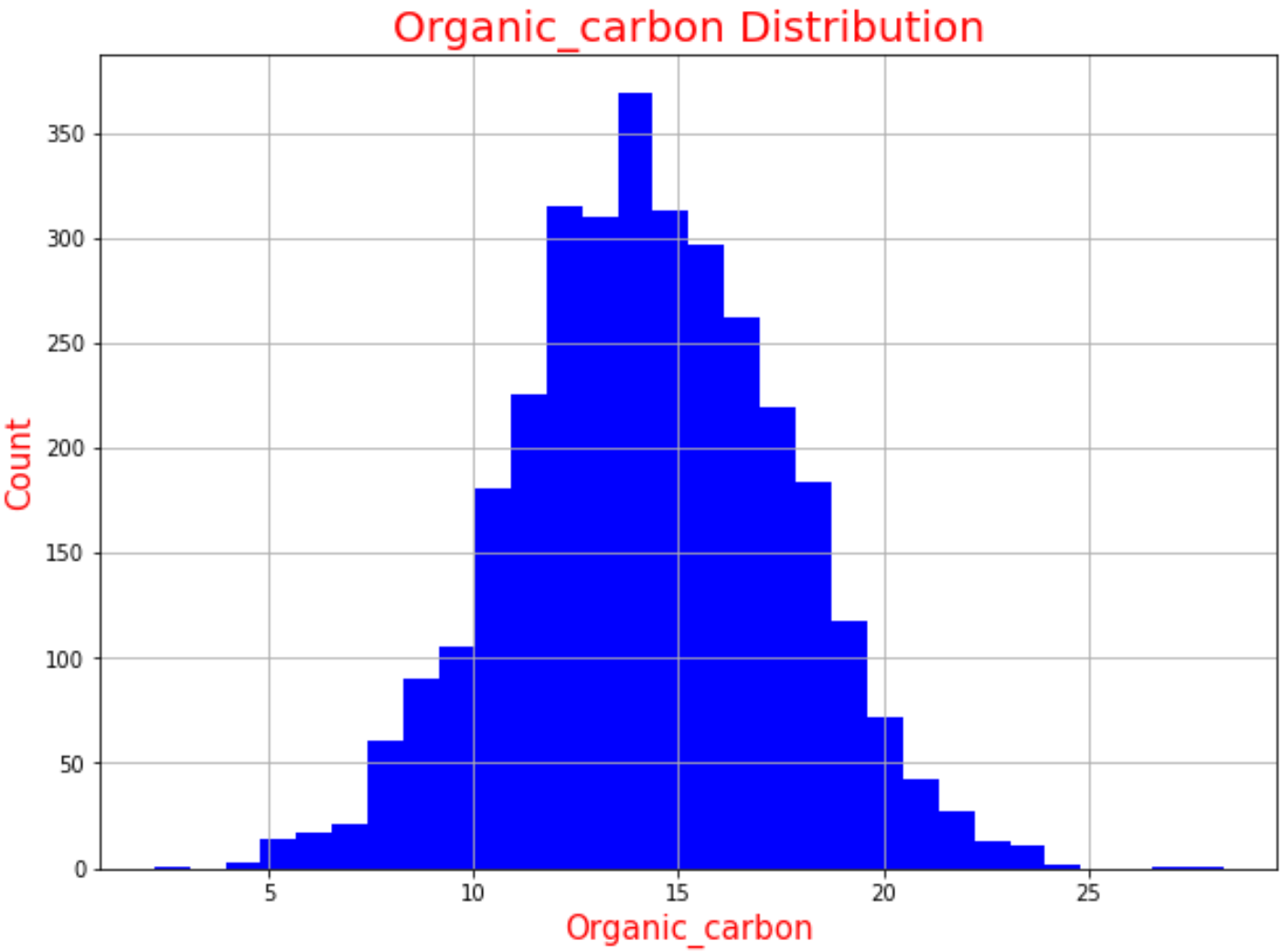

این نمودار نشان دهنده مقدار کلرامین موجود در آب است. همانطور که از نمودار مشخص است بیشترین دیتاهای ثبت شده در بازه ۶ تا ۸ ppm ثبت شده است که با دور شدن از آن از تعداد دیتاهای ثبت شده کاسته میشود.



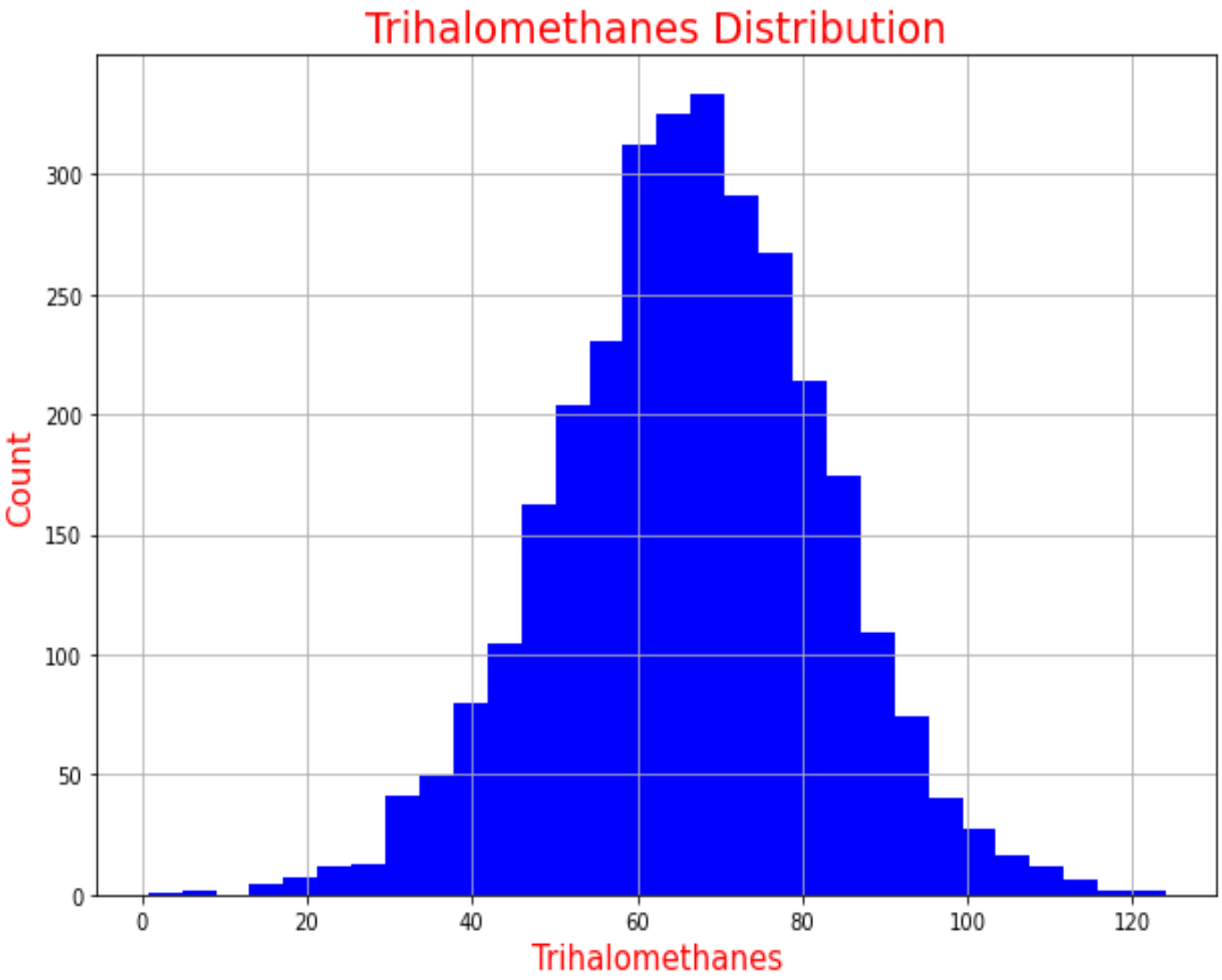

این نمودار نشان دهنده مقدار سولفات های محلول در آب برحسب میلی گرم در لیتر است. همانطور که از شکل مشخص است بیشترین دیتاهای ثبت شده در بازه ۳۰۰ تا ۳۵۰ میلی گرم در لیتر است که به تدریج با افزایش و یا کاهش این مقدار از تعداد دیتاهای ثبت شده کاسته میشود.



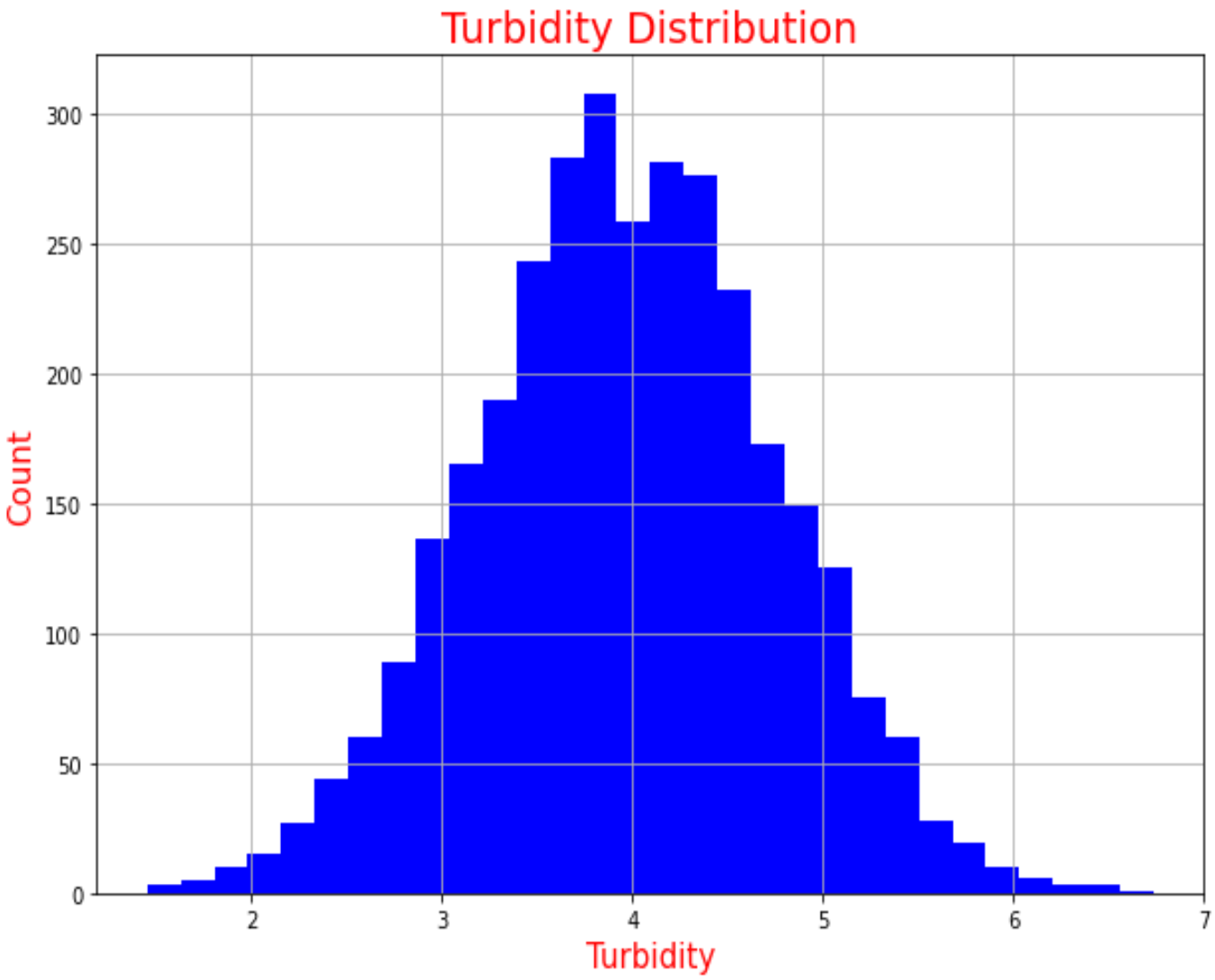

این نمودار نشان دهنده تعداد دیتاهای ثبت شده در هدایت الکتریکی های مختلف می باشد. بیشترین تعداد دیتاها در بازه ۳۵۰ تا ۵۰۰ µS/cm ثبت شده است و به تدریج با فاصله گرفتن از این بازه از تعداد و چگالی دیتاها کاسته میشود.



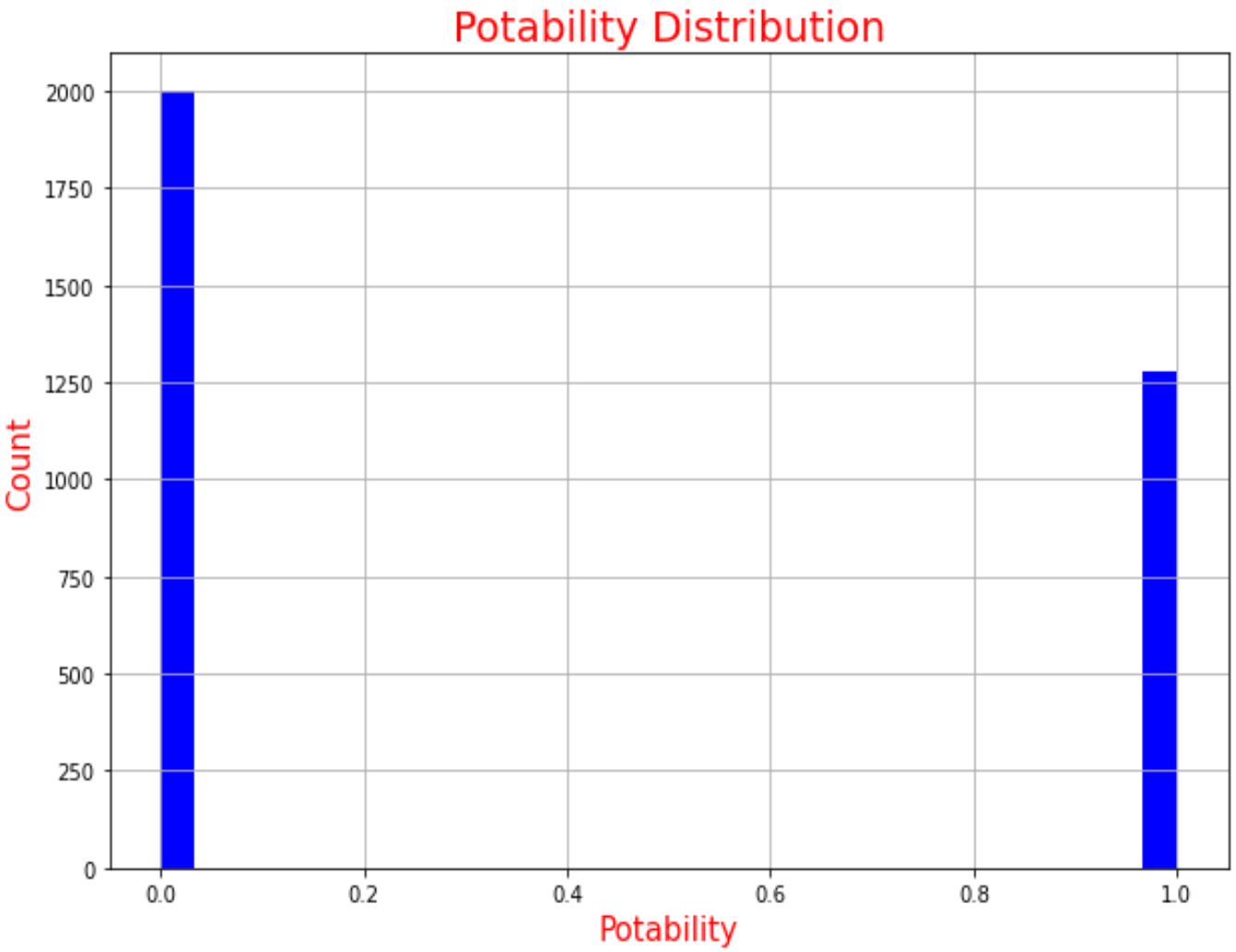

این نمودار نشان دهنده میزان کربن آلی در آب برحسب ppm است. بیشترین دیتاهای ثبت شده در بازه تقریبی ۱۰ تا ۱۷ ppm می باشد و با فاصله گرفتن از این مقدار از تعداد دیتاها کاسته میشود. همچنین بیشترین تعداد دیتا در مقدار تقریبی ۱۴ ppm ثبت شده است.




این نمودار نشان دهنده میزان تری هالومتان ها بر حسب میکروگرم در لیتر در آب است. همانطور که مشخص است بیشترین دیتاهای ثبت شده در بازه ۵۰ تا ۸۰ میکروگرم در لیتر است که با فاصله گرفتن از این مقادیر از تعداد دیتاهای ثبت شده کاسته میشود.

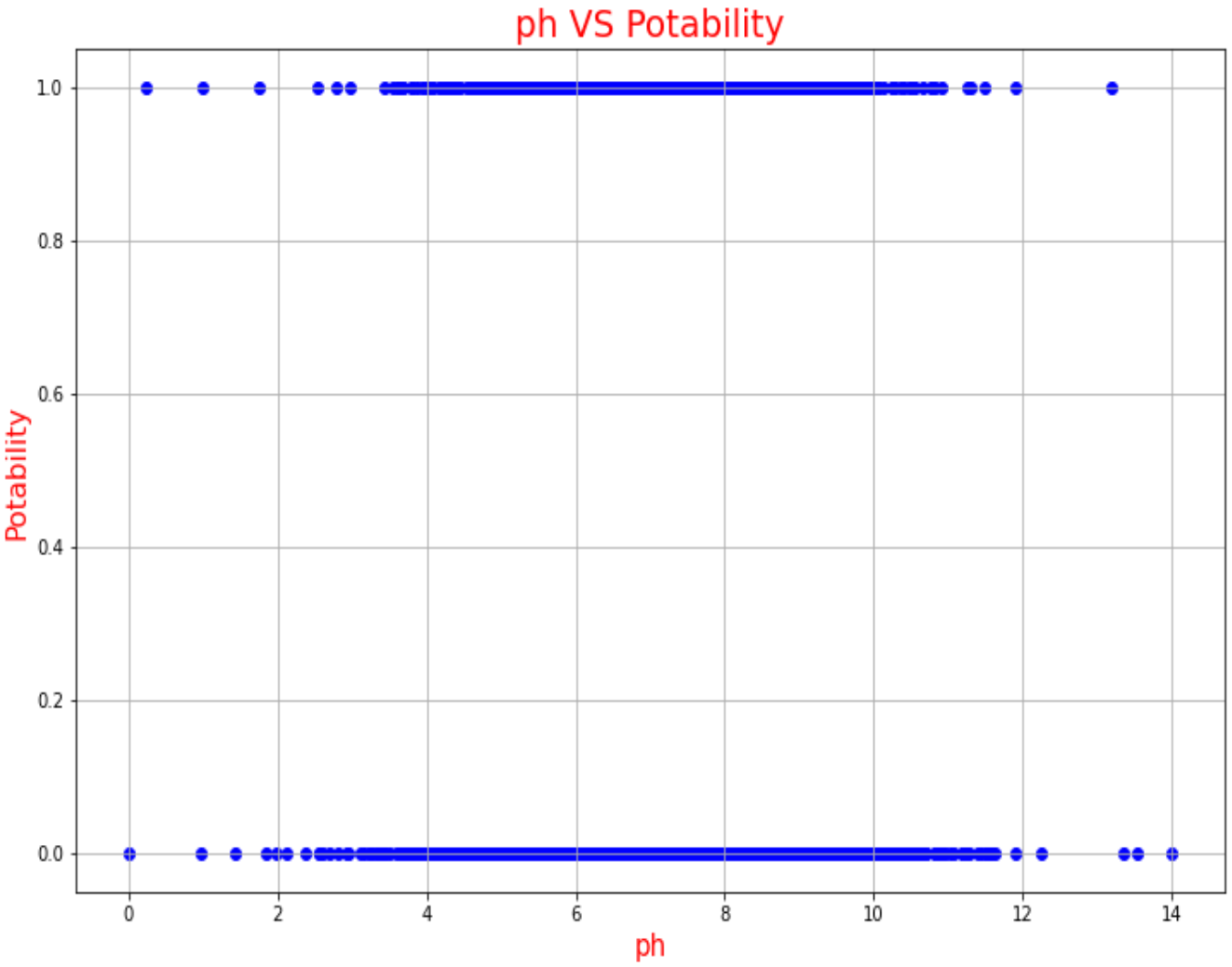



این نمودار میزان خاصیت انعکاس نور در آب را برای مقادیر مختلف نشان میدهد که بیشترین دیتاهای ثبت شده برای آن در بازه ۳ تا ۵ است و با فاصله گرفتن از این بازه از تعداد دیتاها کاسته میشود همچنین همانطور که مشخص است کل دیتاهای خوانده شده در بازه ۰ تا ۷ میباشد.

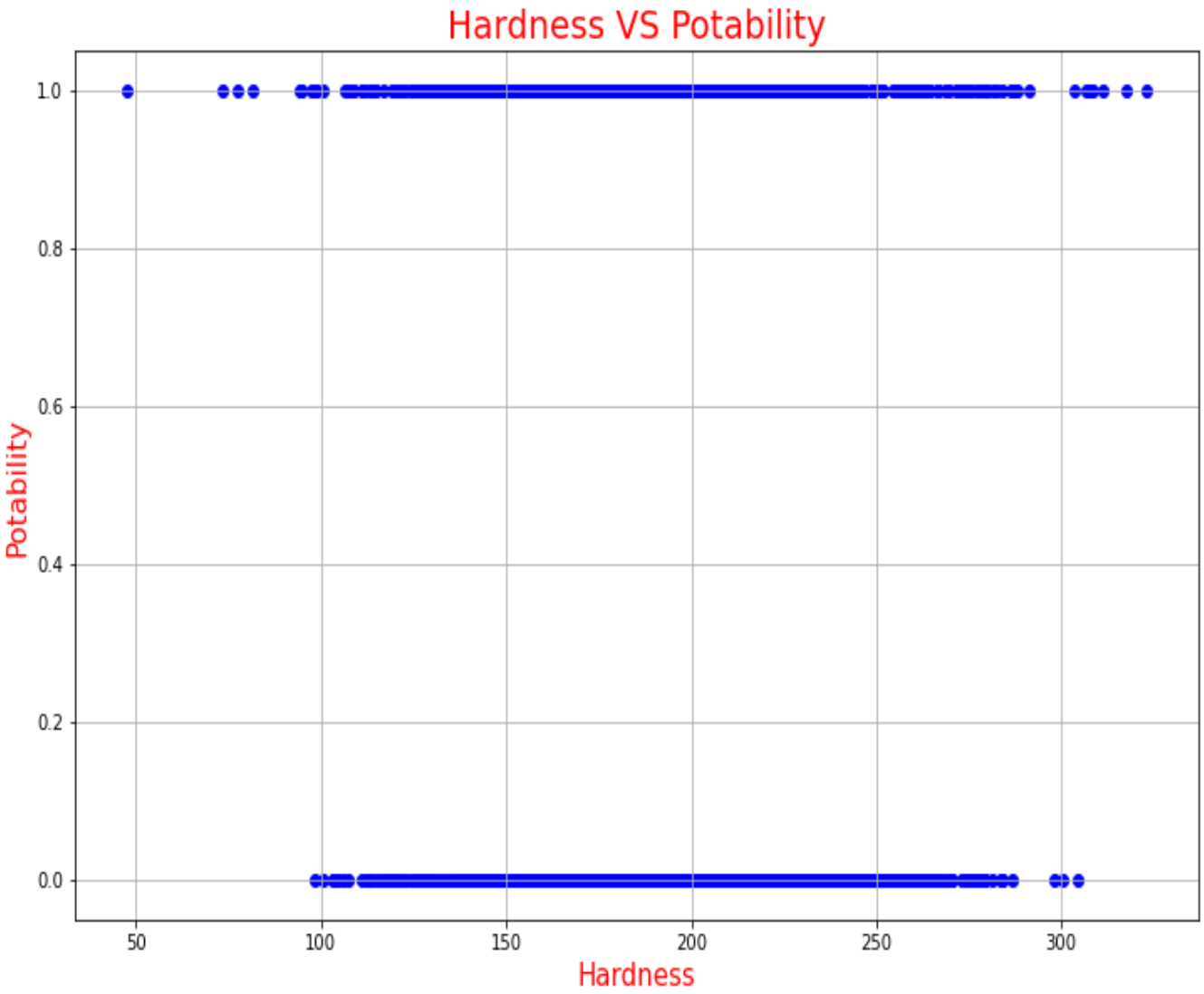



مقدار ۱ در این نمودار بیانگر آب قابل مصرف و مقدار ۰ بیانگر آب غیرقابل مصرف میباشد که مشخصا با بررسی سایر ویژگی های دیتاهای خوانده شده تعداد زیادی از دیتاها قابل مصرف برای استخر نیستند (نزدیک به ۲۰۰۰ دیتا)

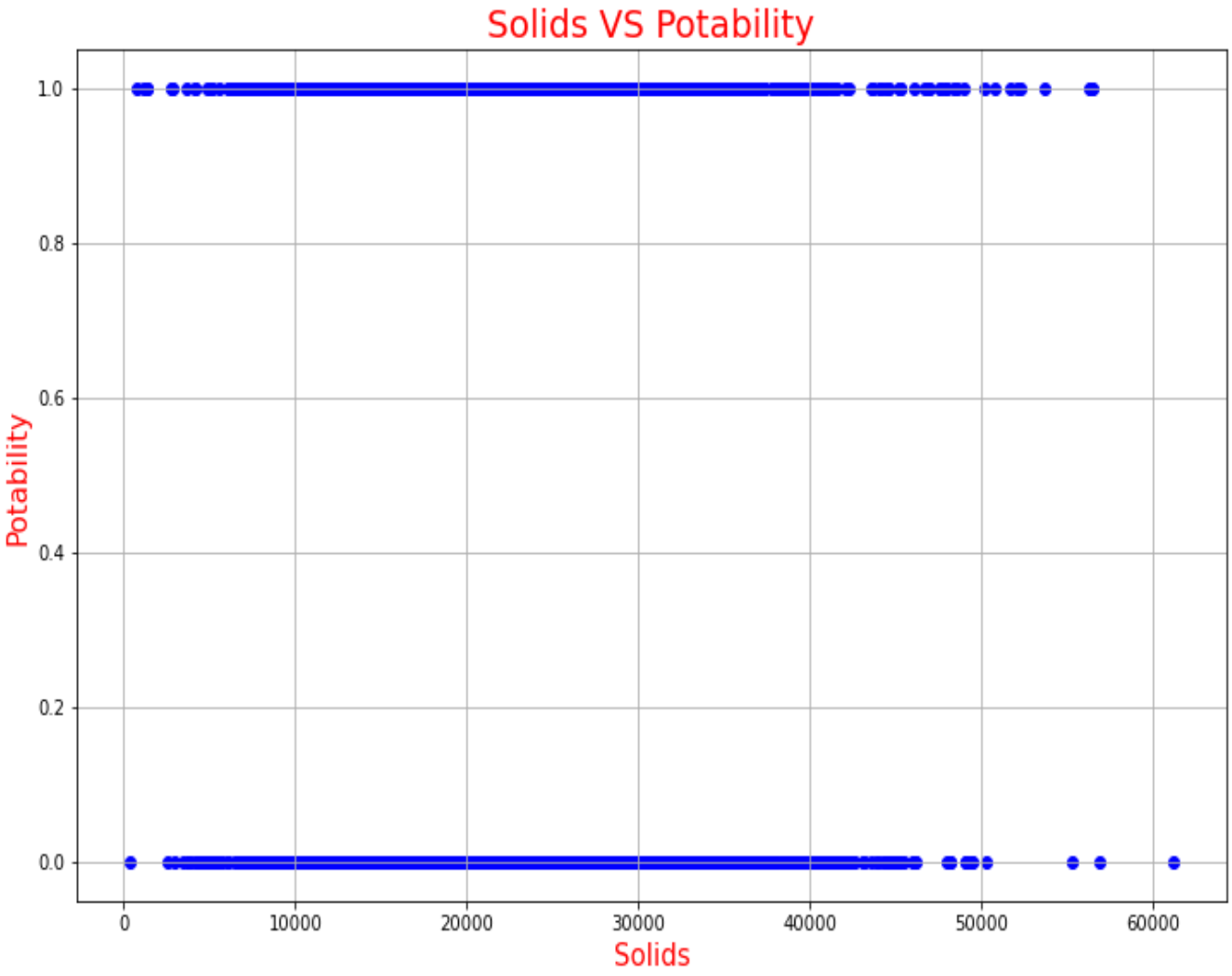
بررسی دو به دوی فیچرها با تارگت :



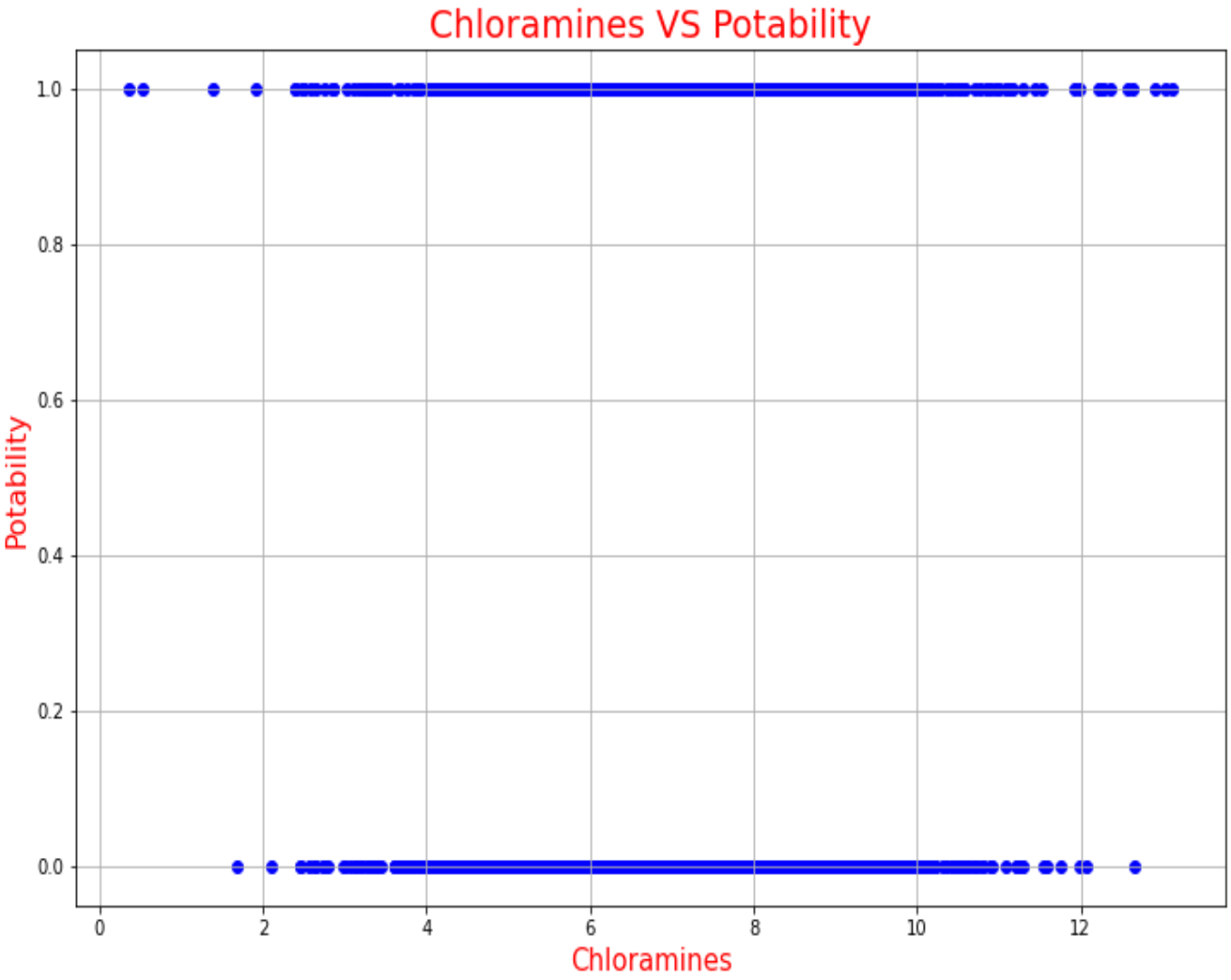

شکل بالا میزان ph آب را براساس قابل مصرف بودن یا نبودن مقایسه میکند بدون درنظر گرفتن دیتاهایی که فاصله بیشتری با چگالی کلی دیتاها دارند می توان گفت که هنگامی که آب قابلیت مصرف دارد بازه ph آن تقریبا بین ۳,۵ تا تا ۱۰,۵ است اما هنگامی که قابلیت مصرف ندارد تقریبا بین ۲ تا ۱۲ است.



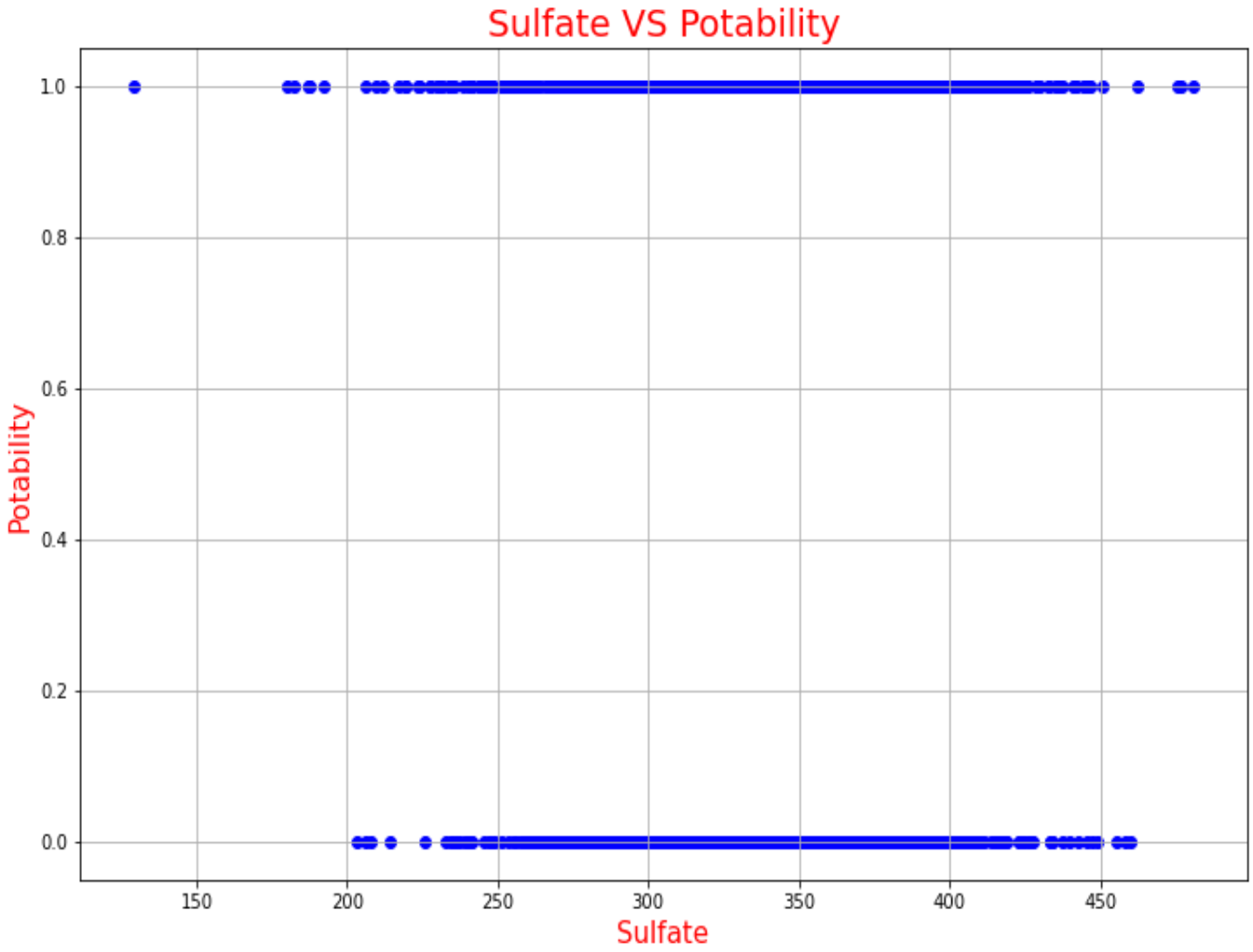

شکل بالا میزان سختی آب را براساس اینکه قابلیت مصرف دارد یا خیر مقایسه کرده است. بازه تغییرات سختی آب در شکل مشخص است. مقادیر کمتر از ۷۵ آب نرم ، بین ۷۵ تا ۲۵۰ آب متوسط و بیشتر از ۲۵۰ آب سخت طبقه بندی میشوند.



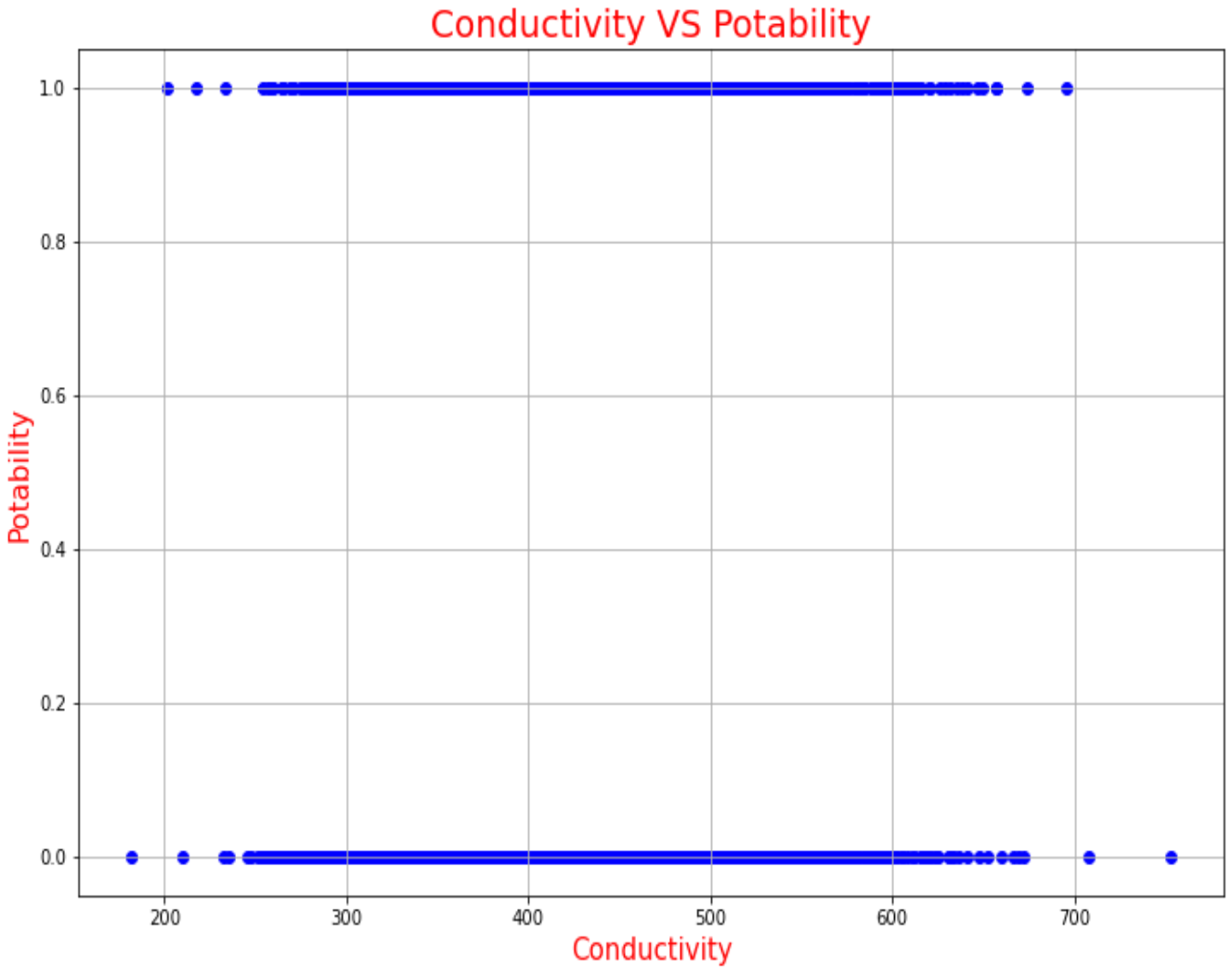

شکل بالا میزان کل مواد جامد محلول در آب را براساس اینکه قابلیت مصرف دارد یا خیر بررسی میکند.



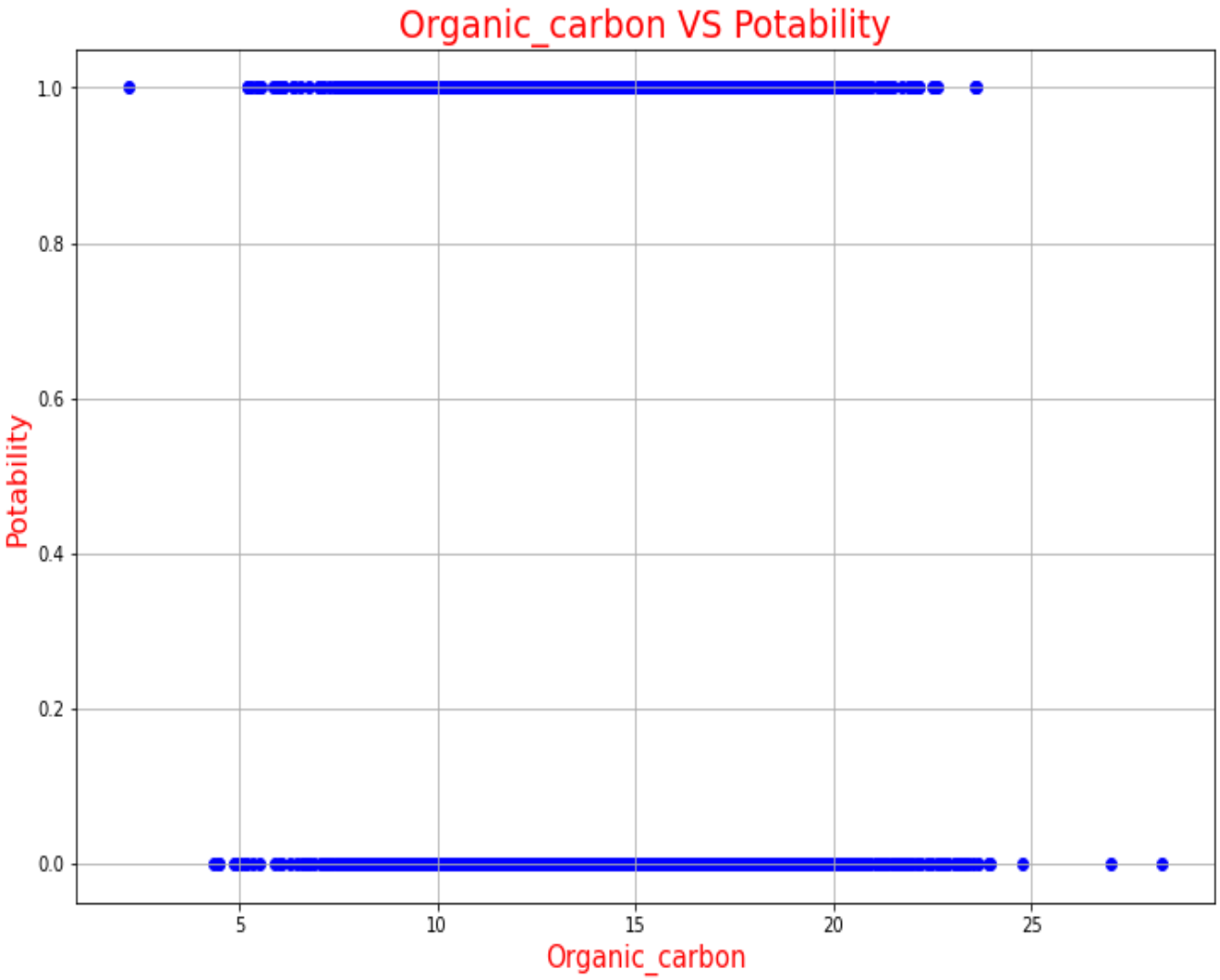

شکل بالا میزان کلرامین در آب را براساس اینکه قابلیت مصرف دارد یا خیر بررسی می کند.



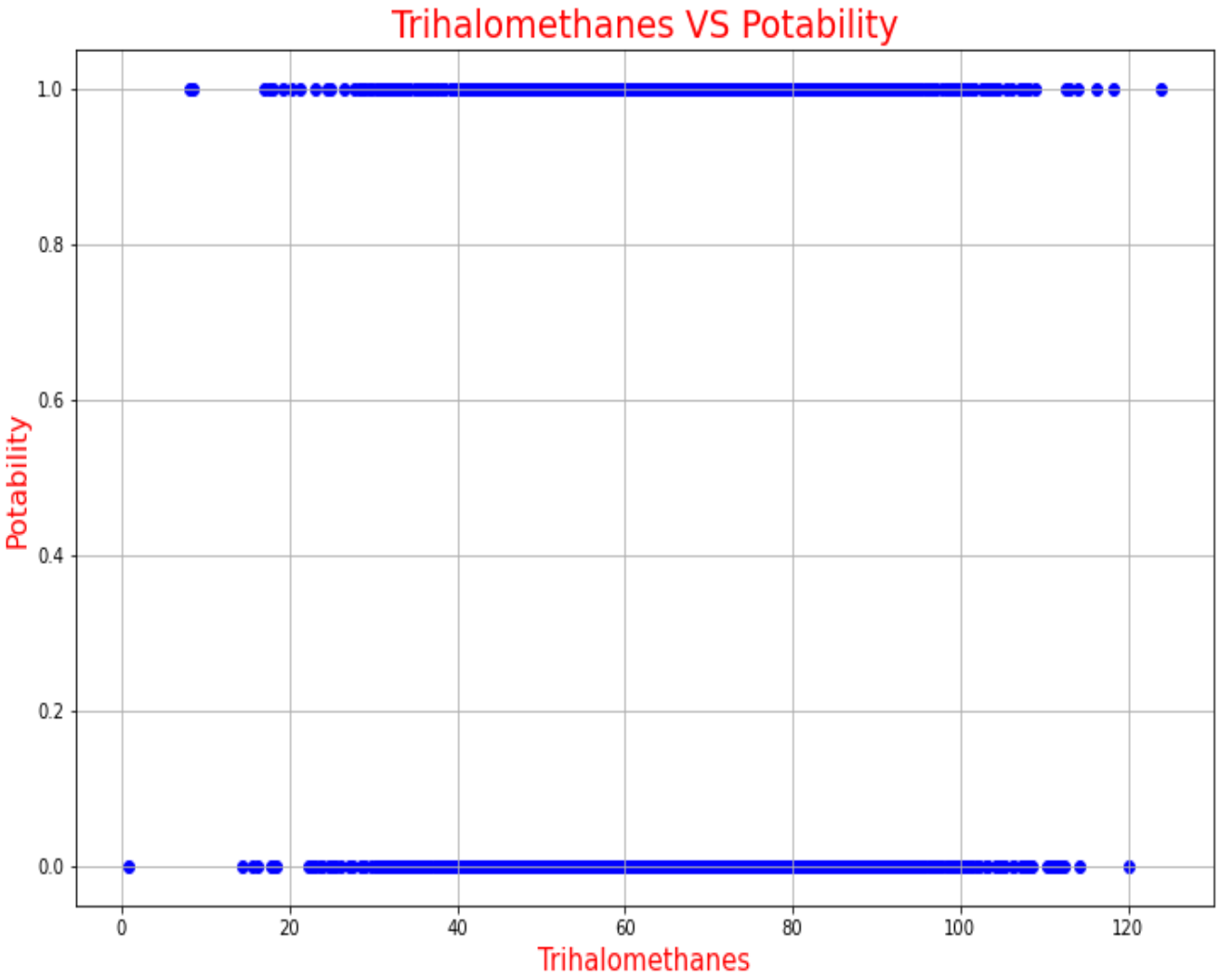

شکل بالا مقدار سولفات های محلول برحسب میلی گرم در لیتر در آب را براساس اینکه آیا آب قابلیت مصرف دارد یا خیر را بررسی میکند.



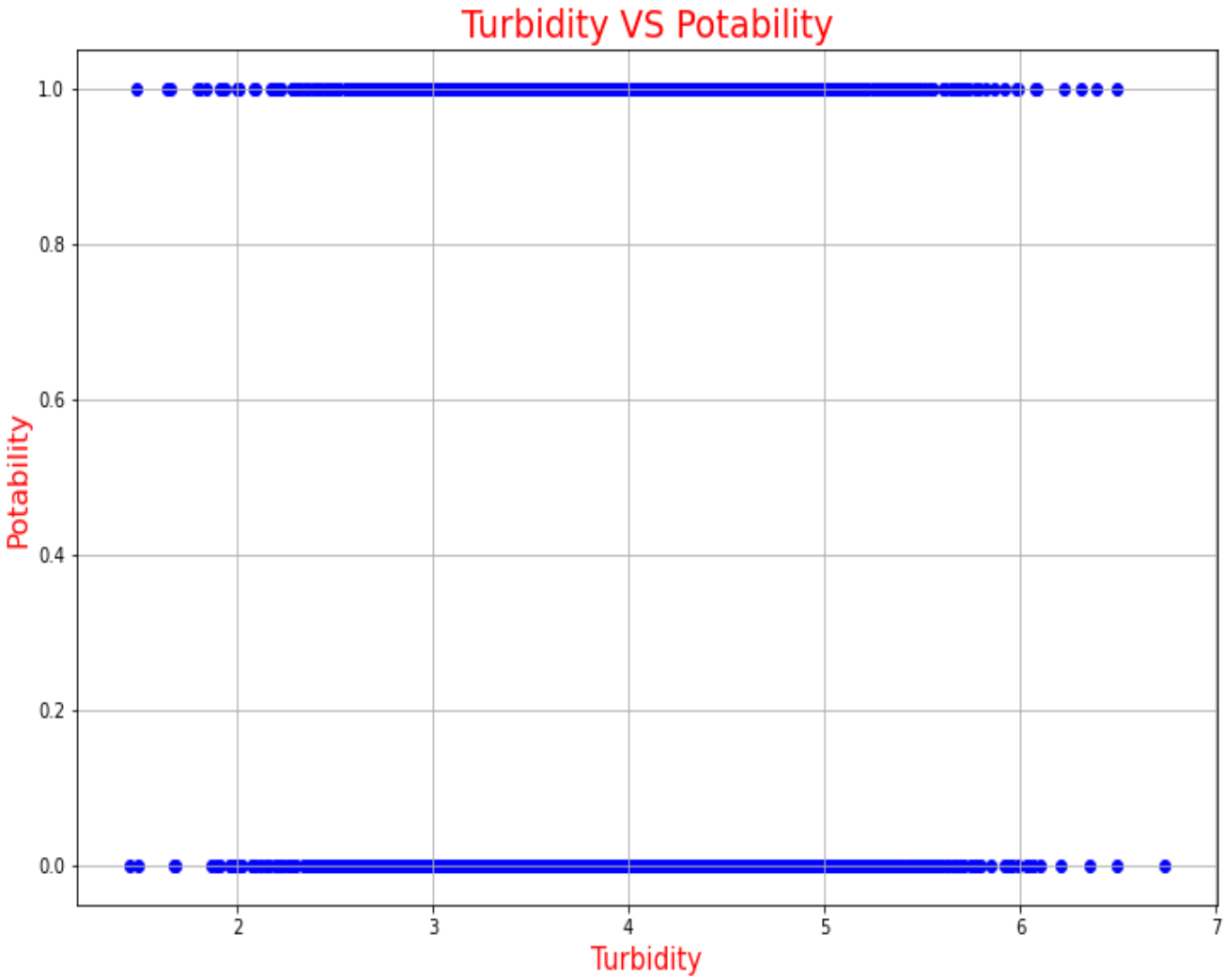

شکل بالا میزان هدایت الکتریکی آب را براساس اینکه آیا قابلیت مصرف دارد یا خیر بررسی میکند.




شکل بالا میزان کربن آلی در آب را براساس اینکه آیا قابلیت مصرف دارد یا خیر بررسی میکند.

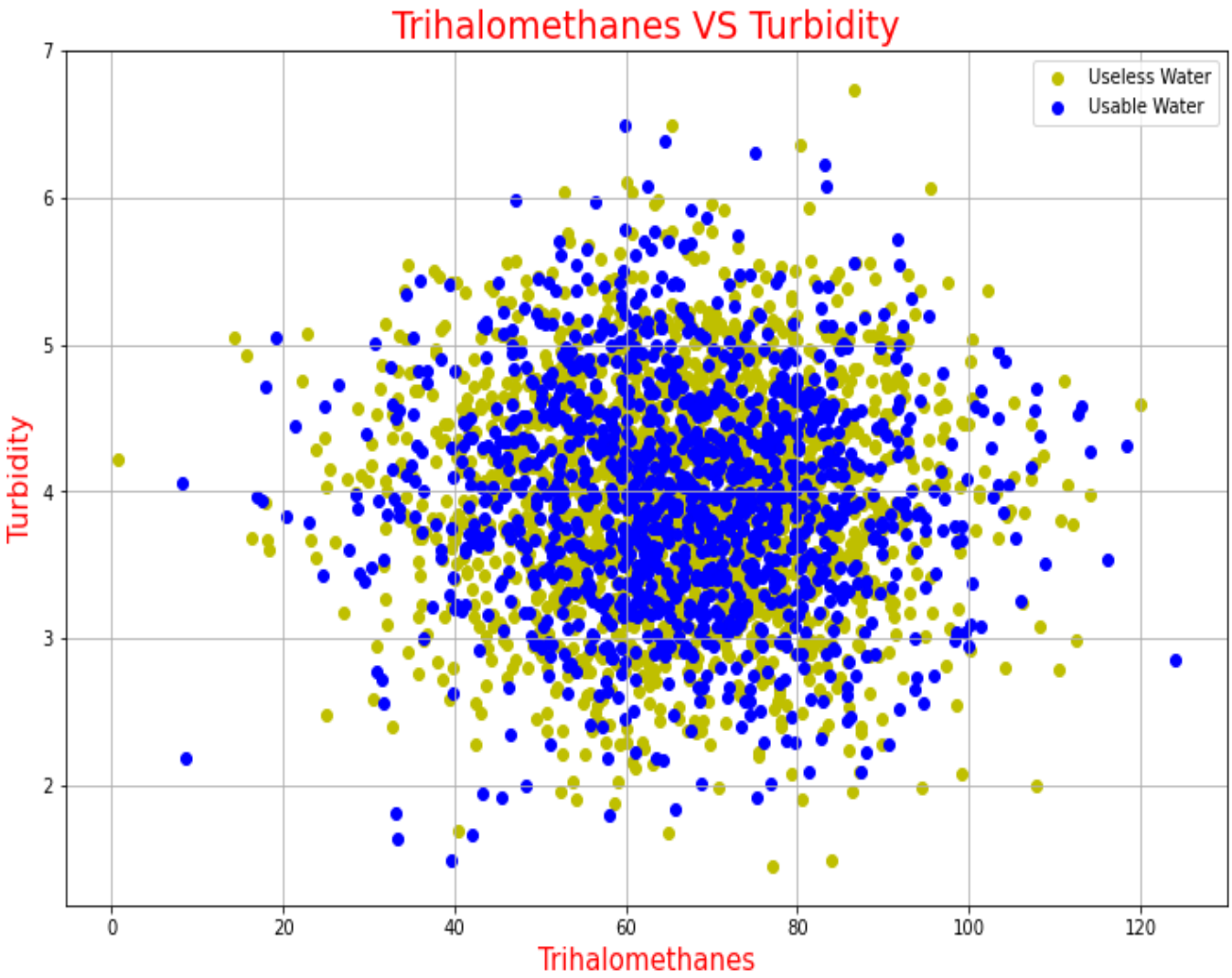



شکل بالا مقدار تری هالومتان ها بر حسب میکروگرم در لیتر در آب را براساس اینکه آیا قابلیت استفاده را دارد یا خیر بررسی میکند.

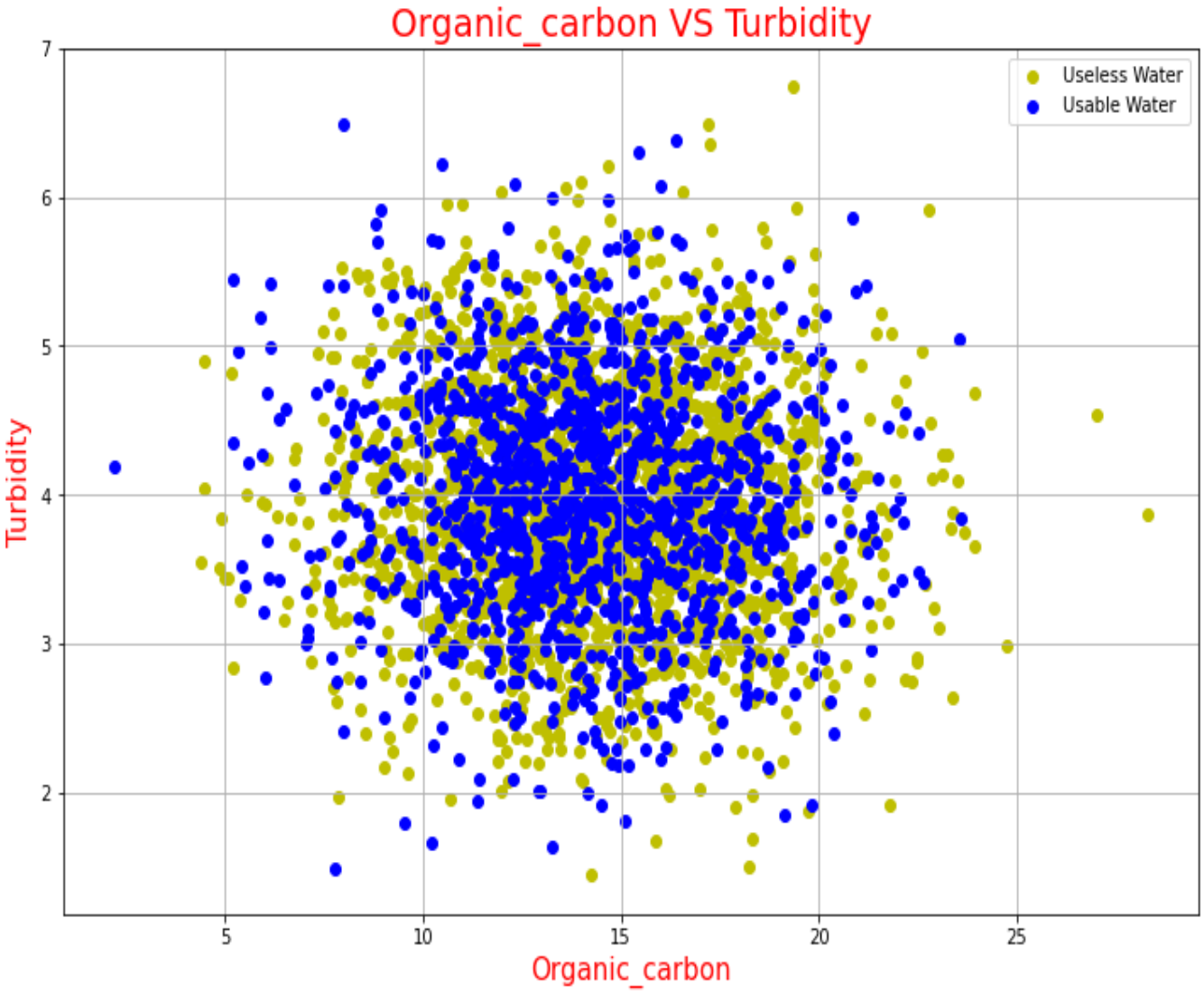



شکل بالا به بررسی قابلیت استفاده آب در خاصیت های مختلف آب در انعکاس نور پرداخته است.

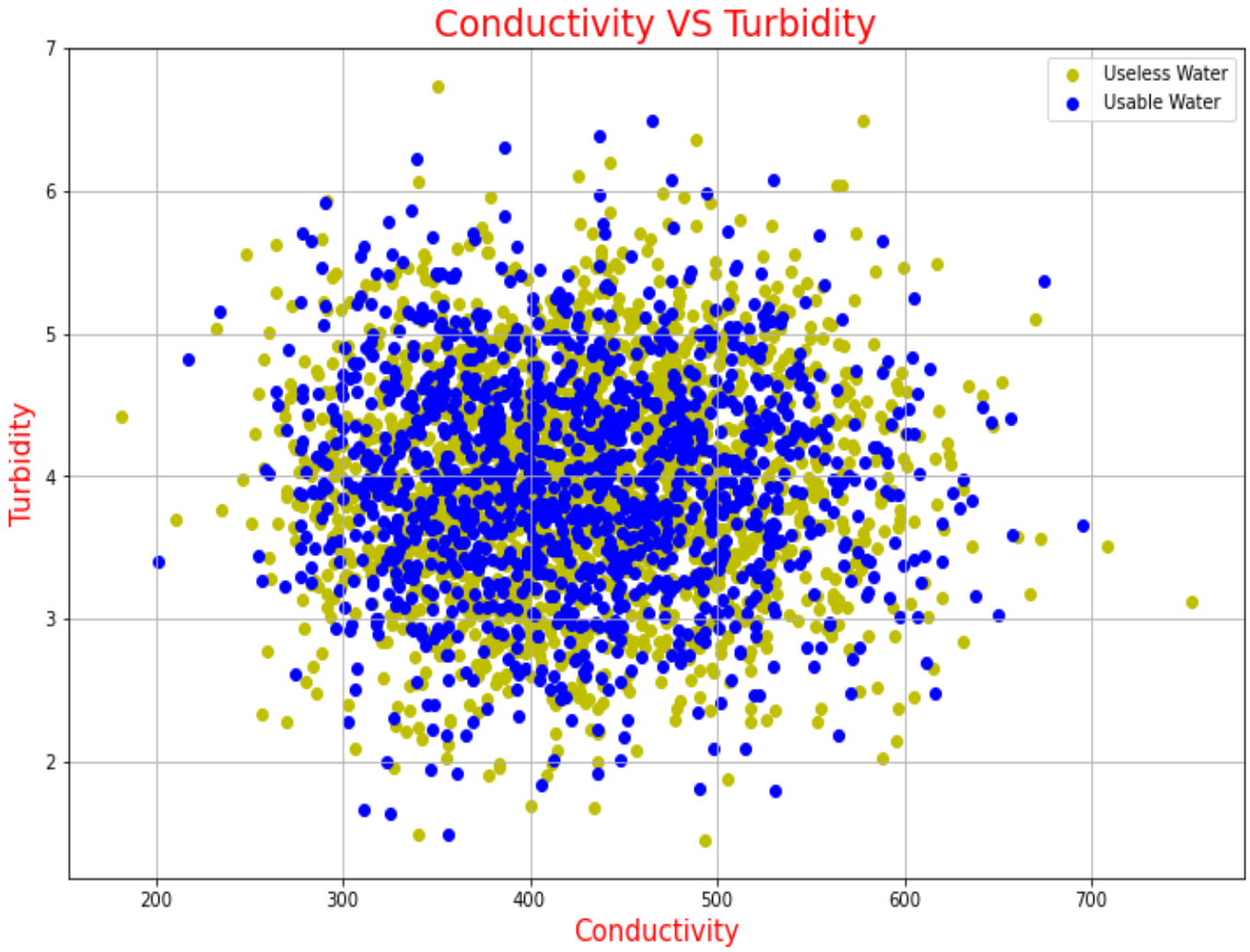
بررسی دو به دوی فیچرها با یکدیگر :



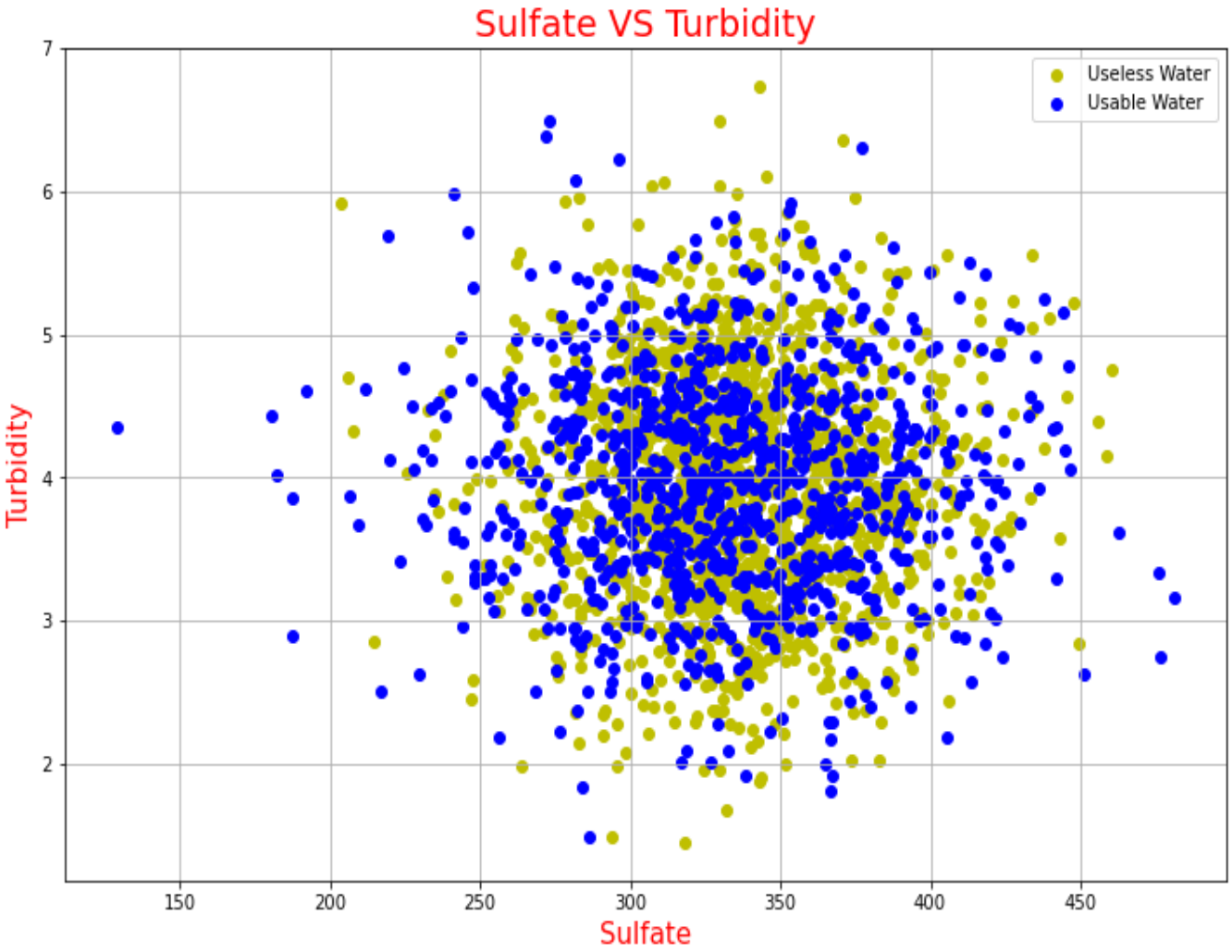

شکل بالا مقدار تری هالومتان ها بر حسب میکروگرم در لیتر را در مقایسه با خاصیت انعکاس نور آب مقایسه میکند. داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در مقدار تری هالومتان ۵۰ تا ۹۰ و خاصیت انعکاس ۳ تا ۵ بیشینه است.



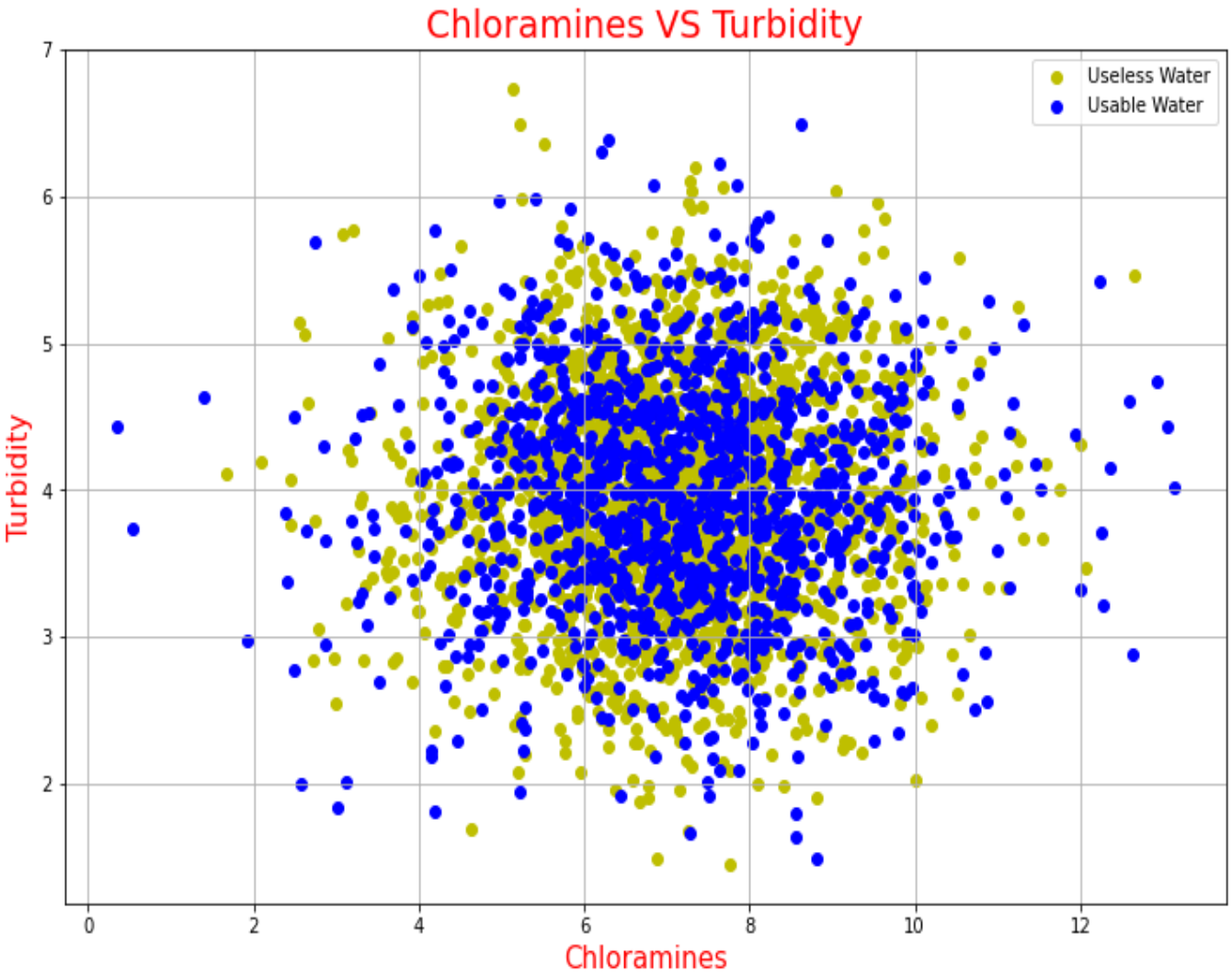

شکل بالا میزان کربن آلی را  در مقایسه با خاصیت انعکاس آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در کربن آلی بین ۱۰ تا ۲۰ و خاصیت انعکاس بین ۳ تا ۵ بیشینه است.



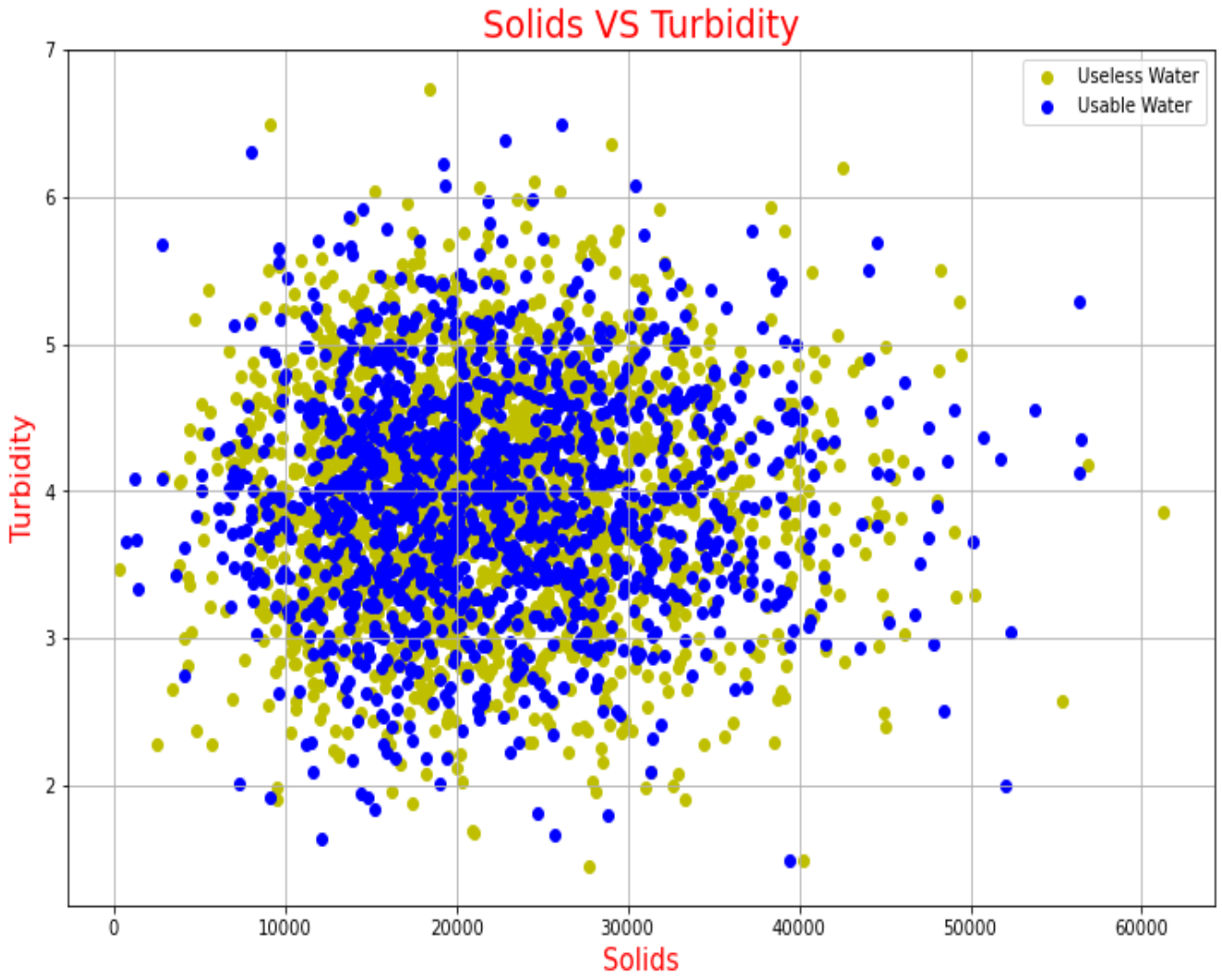

شکل بالا میزان هدایت الکتریکی آب را  در مقایسه با خاصیت انعکاس آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در هدایت الکتریکی بین ۳۰۰ تا ۵۵۰ و خاصیت انعکاس بین ۳ تا ۵ بیشینه است.



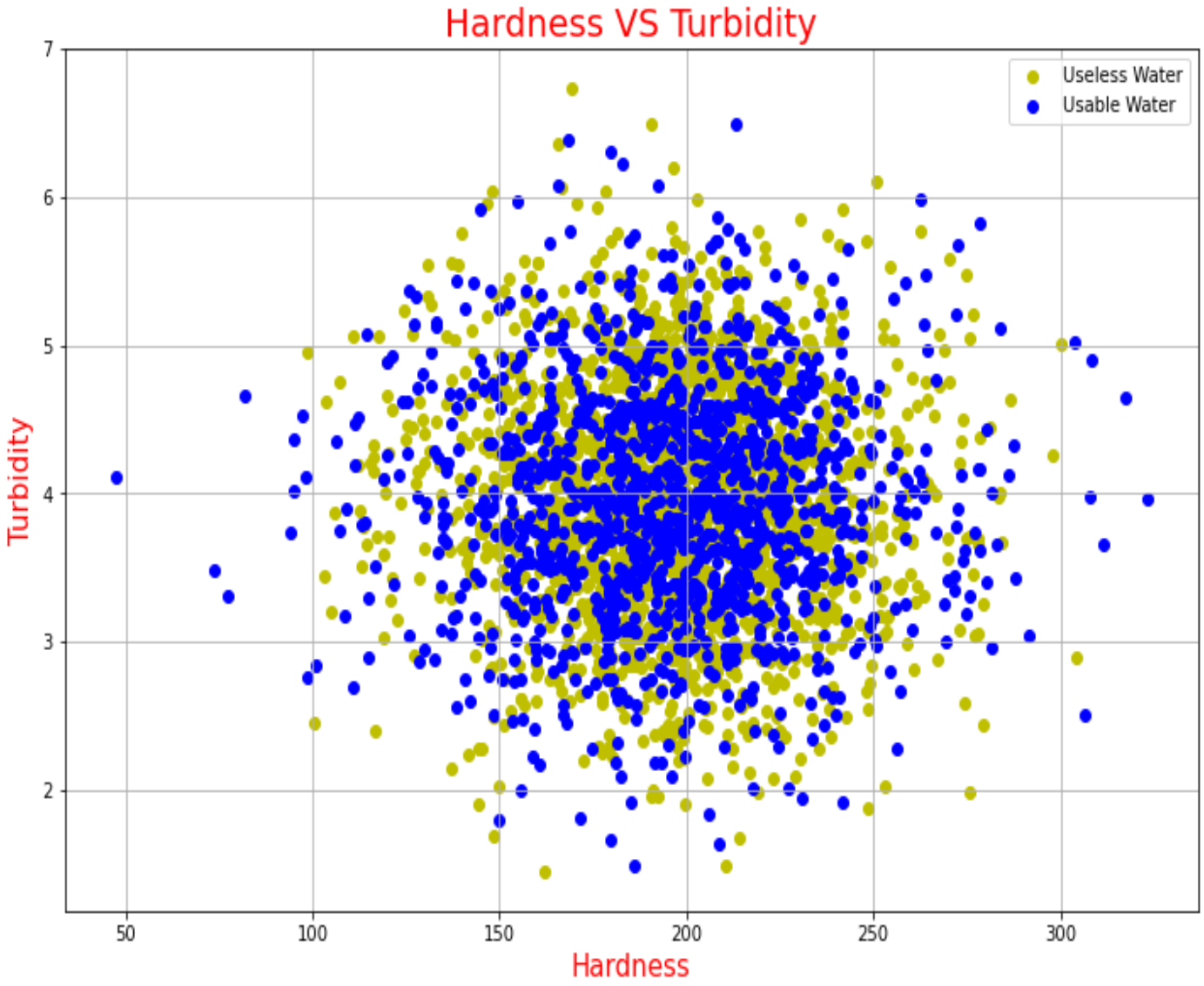

شکل بالا میزان سولفات های محلول برحسب میلی گرم بر لیتر آب را در مقایسه با خاصیت انعکاس آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در محلول های سولفات بین ۳۰۰ تا ۴۰۰ و خاصیت انعکاس بین ۳ تا ۵ بیشینه است.



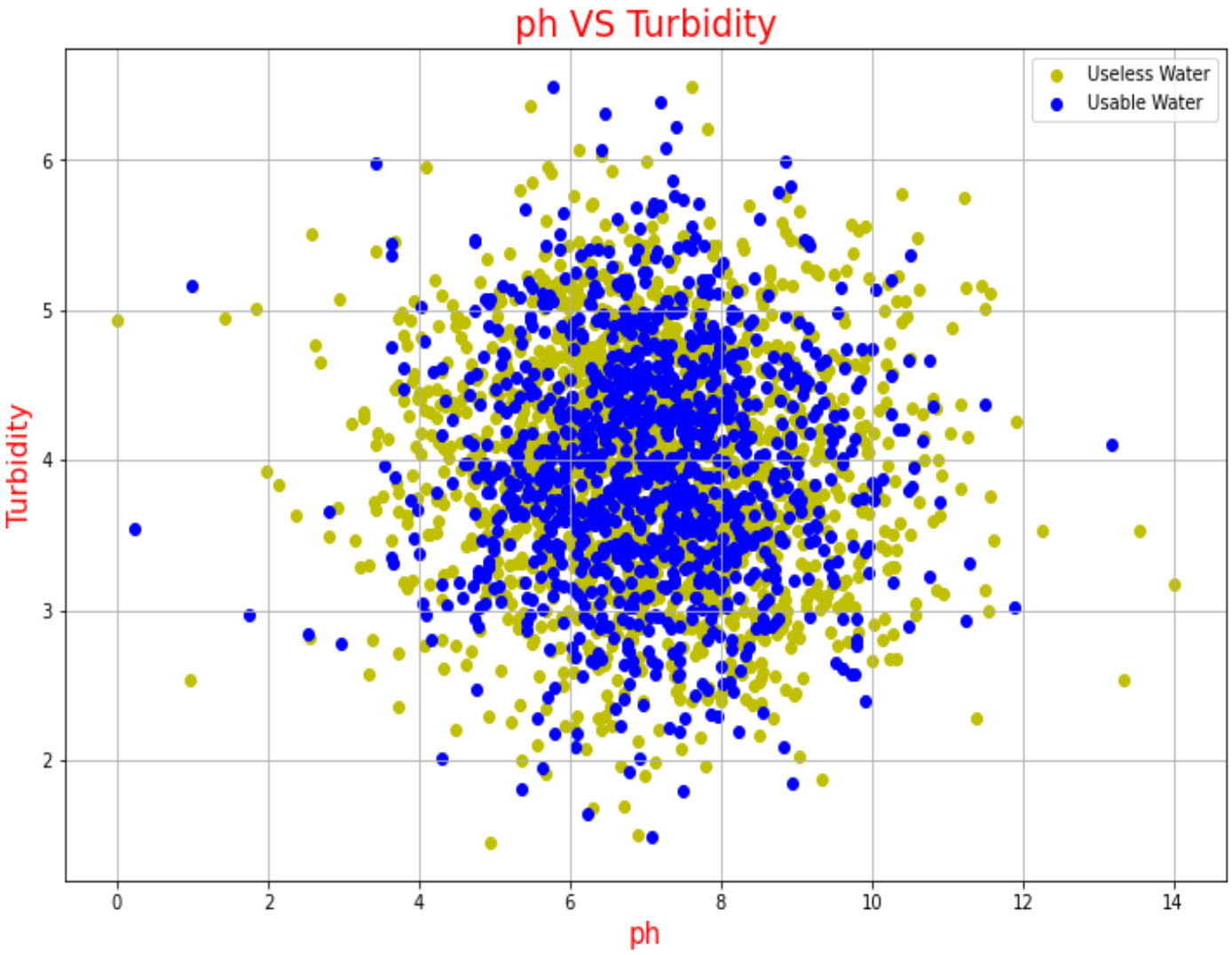

شکل بالا مقدار کلرامین در آب را  در مقایسه با خاصیت انعکاس آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در مقدار کلرامین بین ۵ تا ۹ و خاصیت انعکاس بین ۳ تا ۵ بیشینه است.



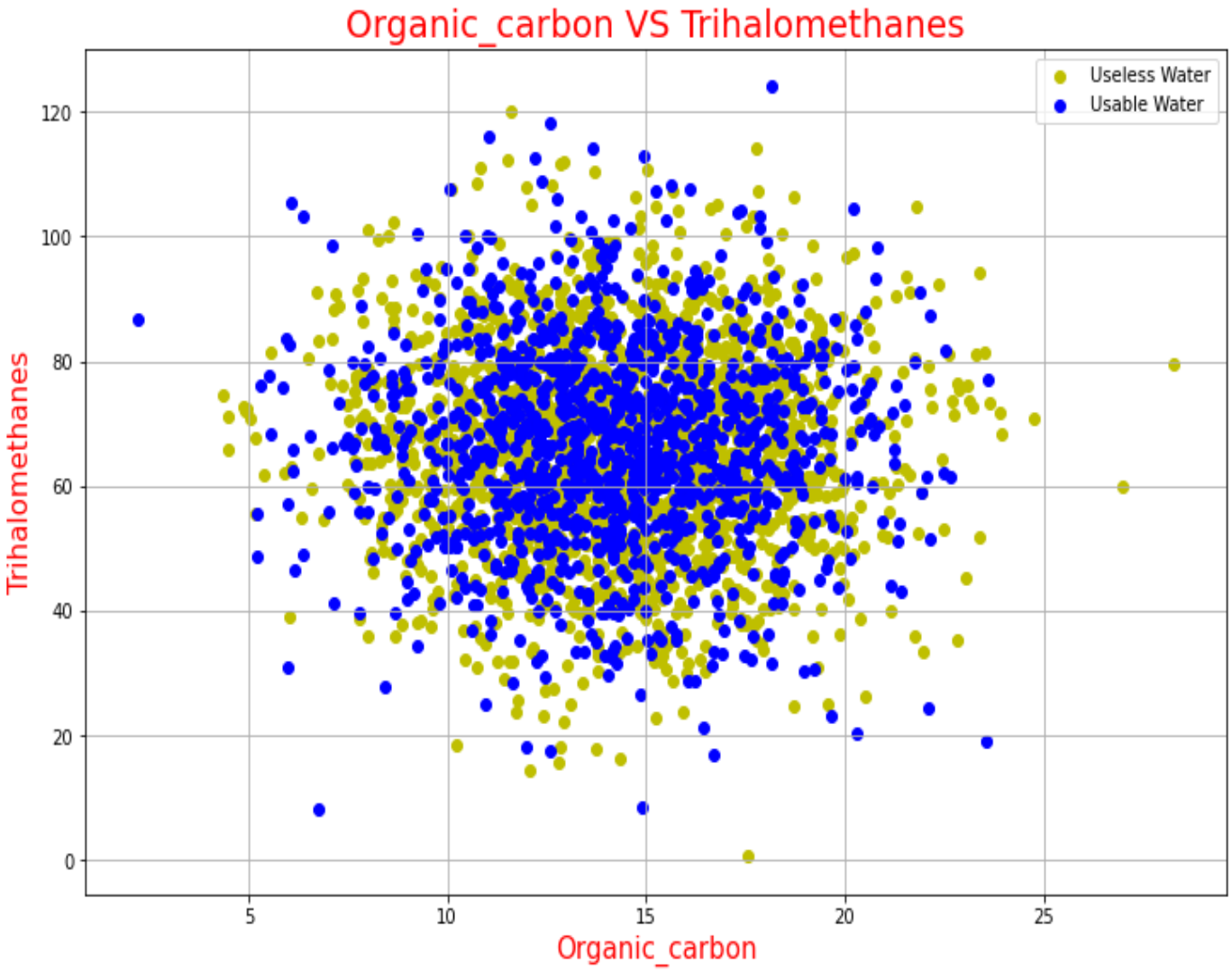

شکل بالا مقدار کل مواد جامد موجود در آب را  در مقایسه با خاصیت انعکاس آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در مقدار مواد جامد بین ۱۰۰۰۰ تا ۳۰۰۰۰ و خاصیت انعکاس بین ۳ تا ۵ بیشینه است.



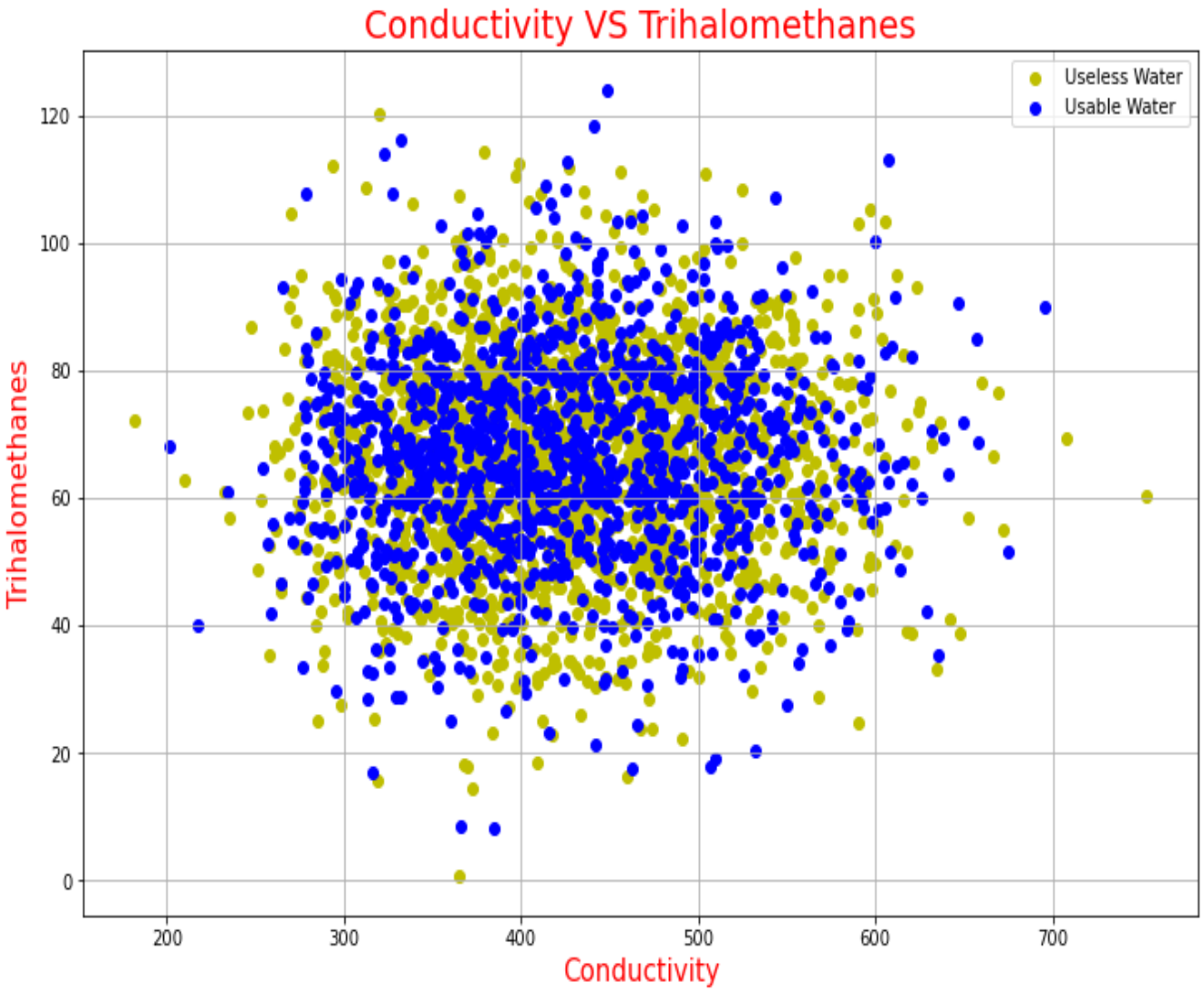

شکل بالا میزان سختی آب را  در مقایسه با خاصیت انعکاس آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در میزان سختی  بین ۱۵۰ تا ۲۵۰ و خاصیت انعکاس بین ۳ تا ۵ بیشینه است.



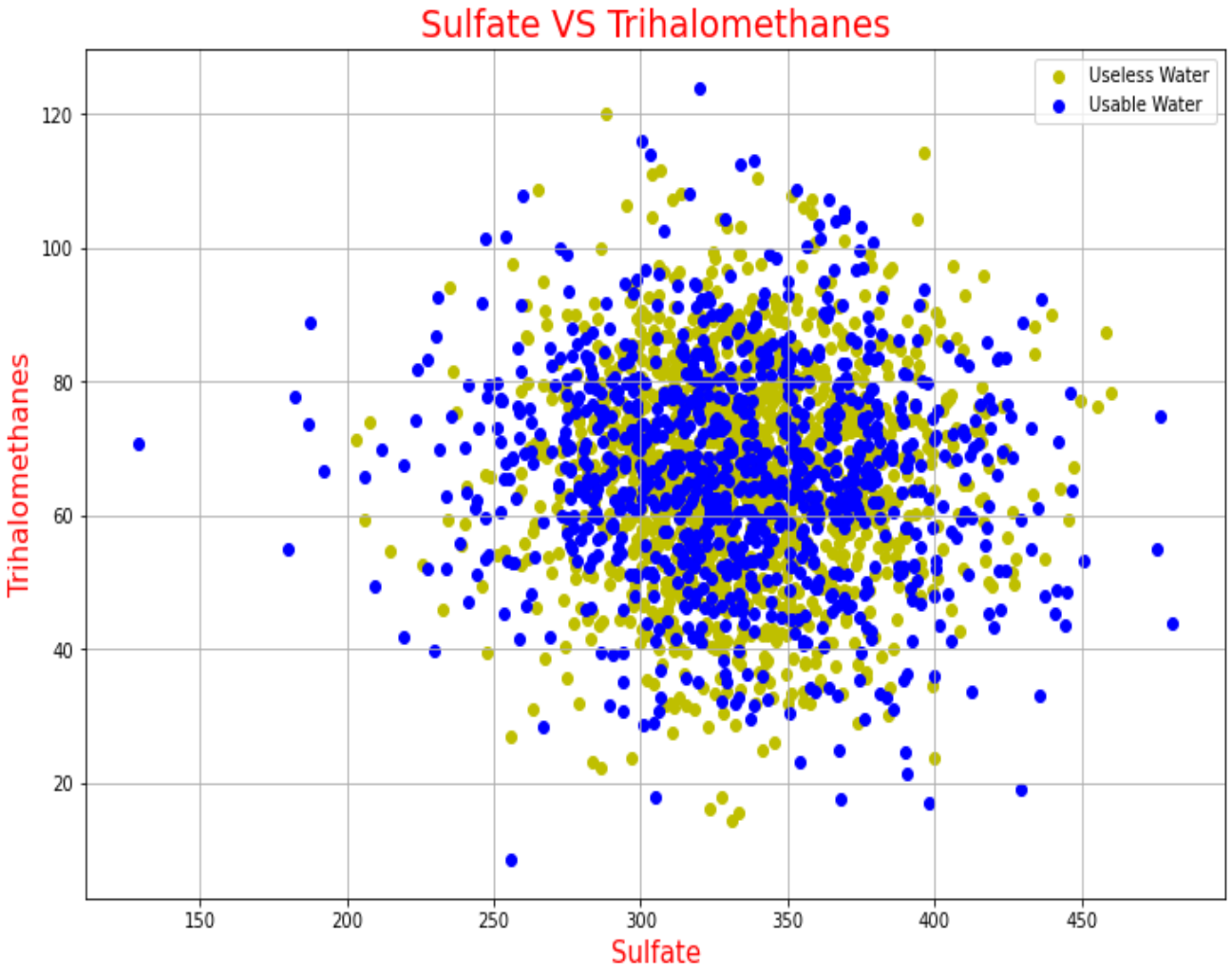

شکل بالا میزان ph آب را در مقایسه با خاصیت انعکاس آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در مقدار ph بین ۴ تا ۱۰ و خاصیت انعکاس بین ۳ تا ۵ بیشینه است همچنین با پیشروی در نمودار به سمت ph های بیشتر یا کمتر از این بازه به تعداد دیتاهای آب غیرقابل استفاده افزوده و از تعداد دیتاهای آب قابل استفاده کاسته میشود.



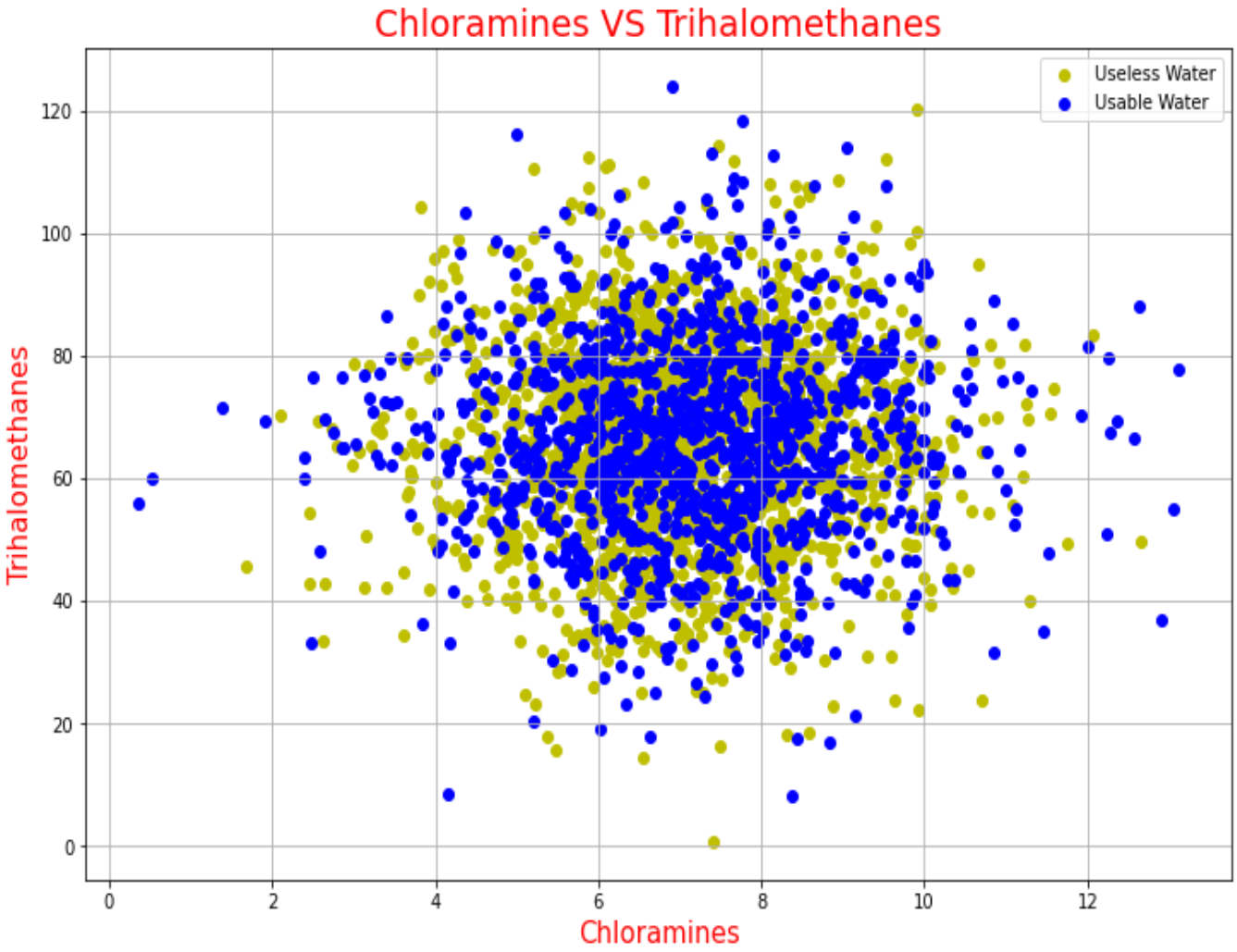

شکل بالا میزان کربن آلی را  در مقایسه با مقدار تری هالومتان ها در آب بر حسب میکروگرم در لیتر نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در مقدار کربن بین ۱۰ تا ۲۰ و مقدار تری هالومتان بین ۵۰ تا ۹۰ بیشینه است. همچنین چگالی دیتاهای آب قابل استفاده در مقدار تری هالومتان بین ۵۰ تا ۹۰ نسبت به چگالی دیتاهای آب غیرقابل استفاده در همین بازه بیشتر است.



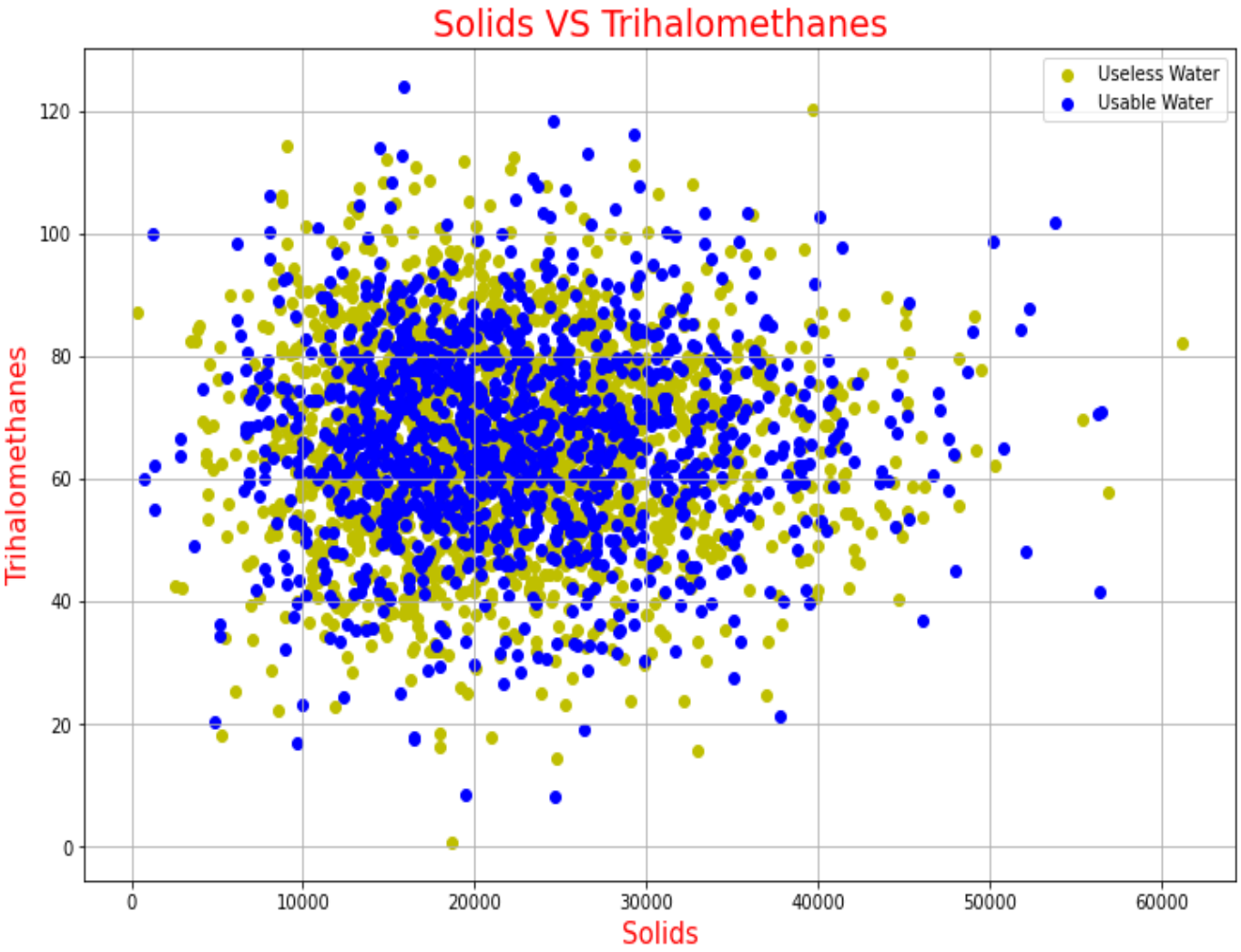

شکل بالا میزان هدایت الکتریکی را در مقایسه با مقدار تری هالومتان ها در آب بر حسب میکروگرم در لیتر نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در میزان هدایت الکتریکی بین ۳۰۰ تا ۵۵۰ و مقدار تری هالومتان بین ۵۰ تا ۹۰ بیشینه است. همچنین چگالی دیتاهای آب قابل استفاده در مقدار تری هالومتان بین ۵۰ تا ۹۰ نسبت به چگالی دیتاهای آب غیرقابل استفاده در همین بازه بیشتر است.



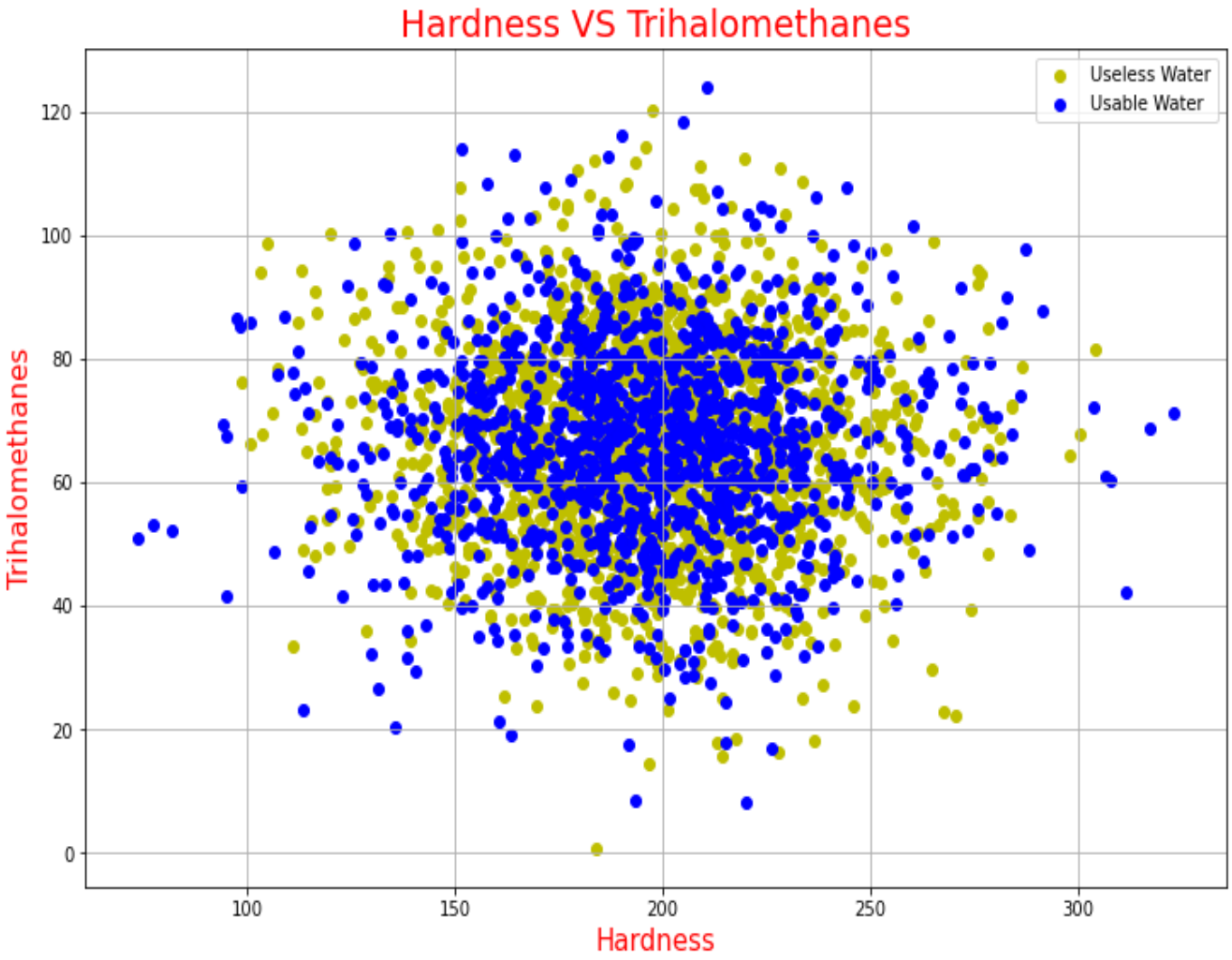

شکل بالا مقدار سولفات های محلول برحسب میلی گرم بر لیتر را در مقایسه با مقدار تری هالومتان ها در آب بر حسب میکروگرم در لیتر نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در مقدار محلول سولفات بین ۳۰۰ تا ۴۰۰ و مقدار تری هالومتان بین ۵۰ تا ۹۰ بیشینه است.



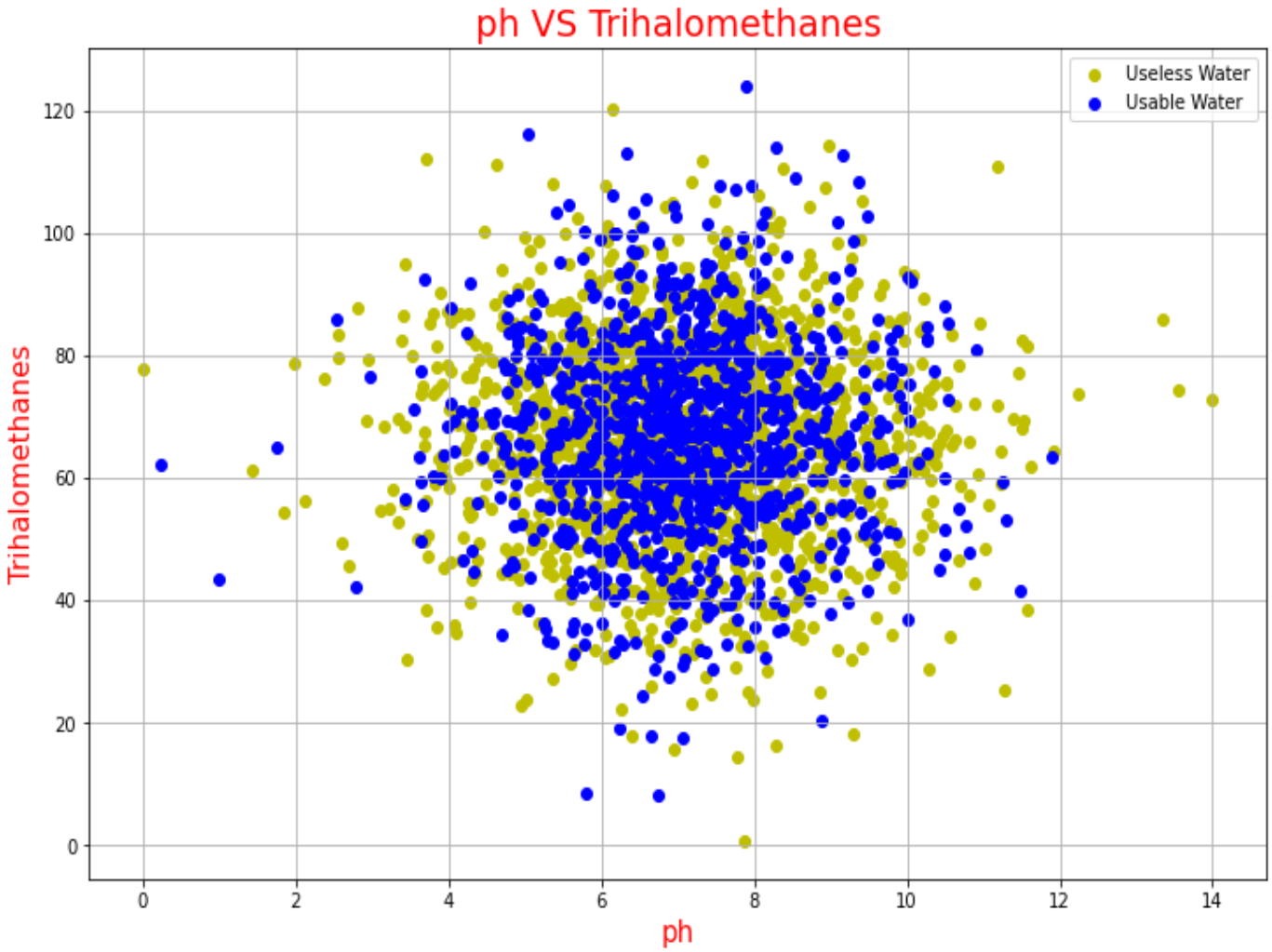

شکل بالا مقدار کلرامین موجود در آب را  در مقایسه با مقدار تری هالومتان ها در آب بر حسب میکروگرم در لیتر نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در مقدار کلرامین  بین ۵ تا ۹ و مقدار تری هالومتان بین ۵۰ تا ۹۰ بیشینه است.



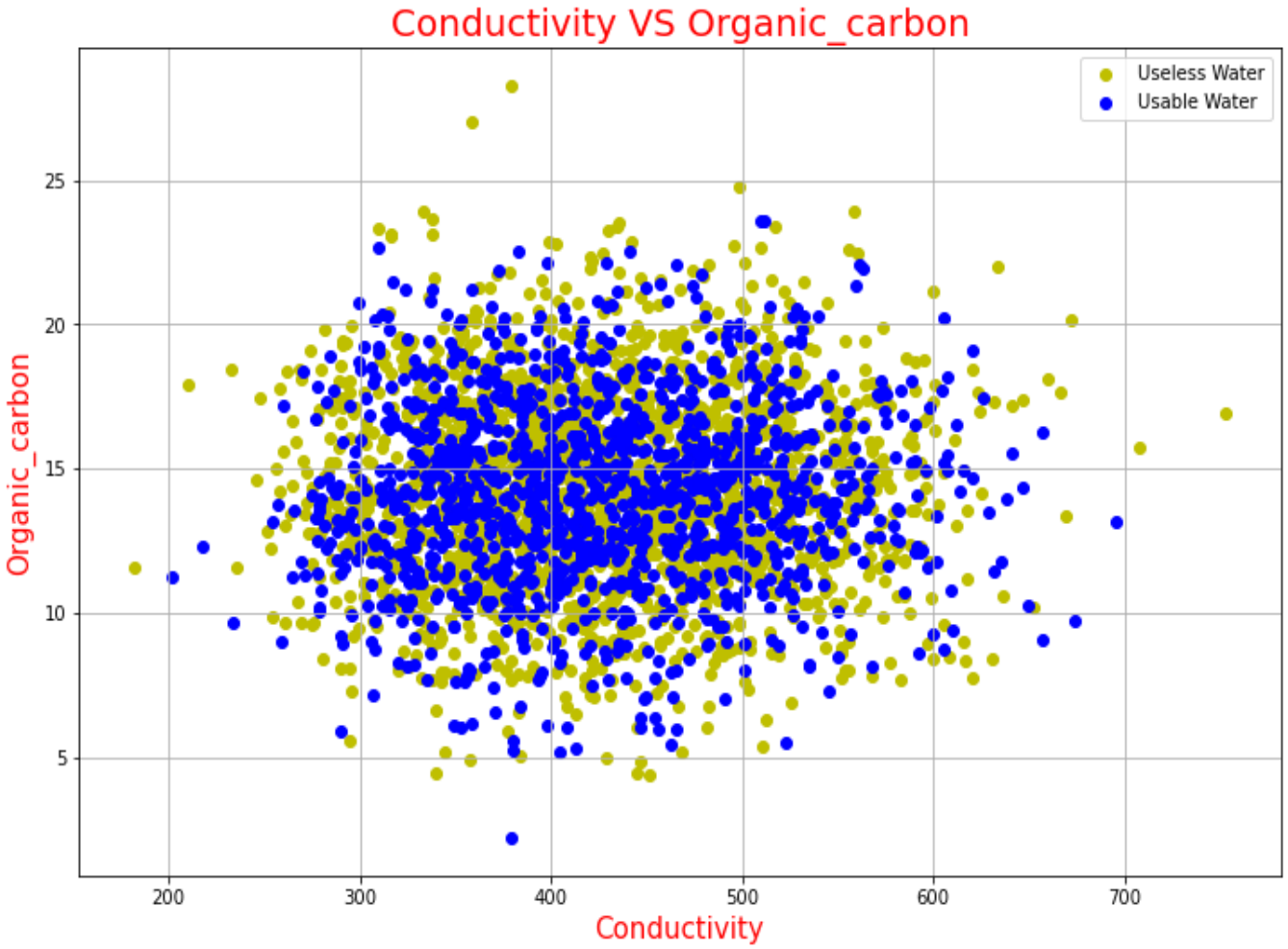

شکل بالا مقدار مواد جامد در آب را  در مقایسه با مقدار تری هالومتان ها در آب بر حسب میکروگرم در لیتر نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در مقدار کل مواد جامد بین ۱۰۰۰۰ تا ۳۰۰۰۰ و مقدار تری هالومتان بین ۵۰ تا ۹۰ بیشینه است. همچنین چگالی دیتاهای آب قابل استفاده در مقدار تری هالومتان بین ۵۰ تا ۹۰ نسبت به چگالی دیتاهای آب غیرقابل استفاده در همین بازه بیشتر است.




شکل بالا میزان سختی آب را  در مقایسه با مقدار تری هالومتان ها در آب بر حسب میکروگرم در لیتر نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در میزان سختی آب بین ۱۵۰ تا ۲۵۰ و مقدار تری هالومتان بین ۵۰ تا ۹۰ بیشینه است. همچنین چگالی دیتاهای آب قابل استفاده در مقدار تری هالومتان بین ۵۰ تا ۹۰ نسبت به چگالی دیتاهای آب غیرقابل استفاده در همین بازه بیشتر است.

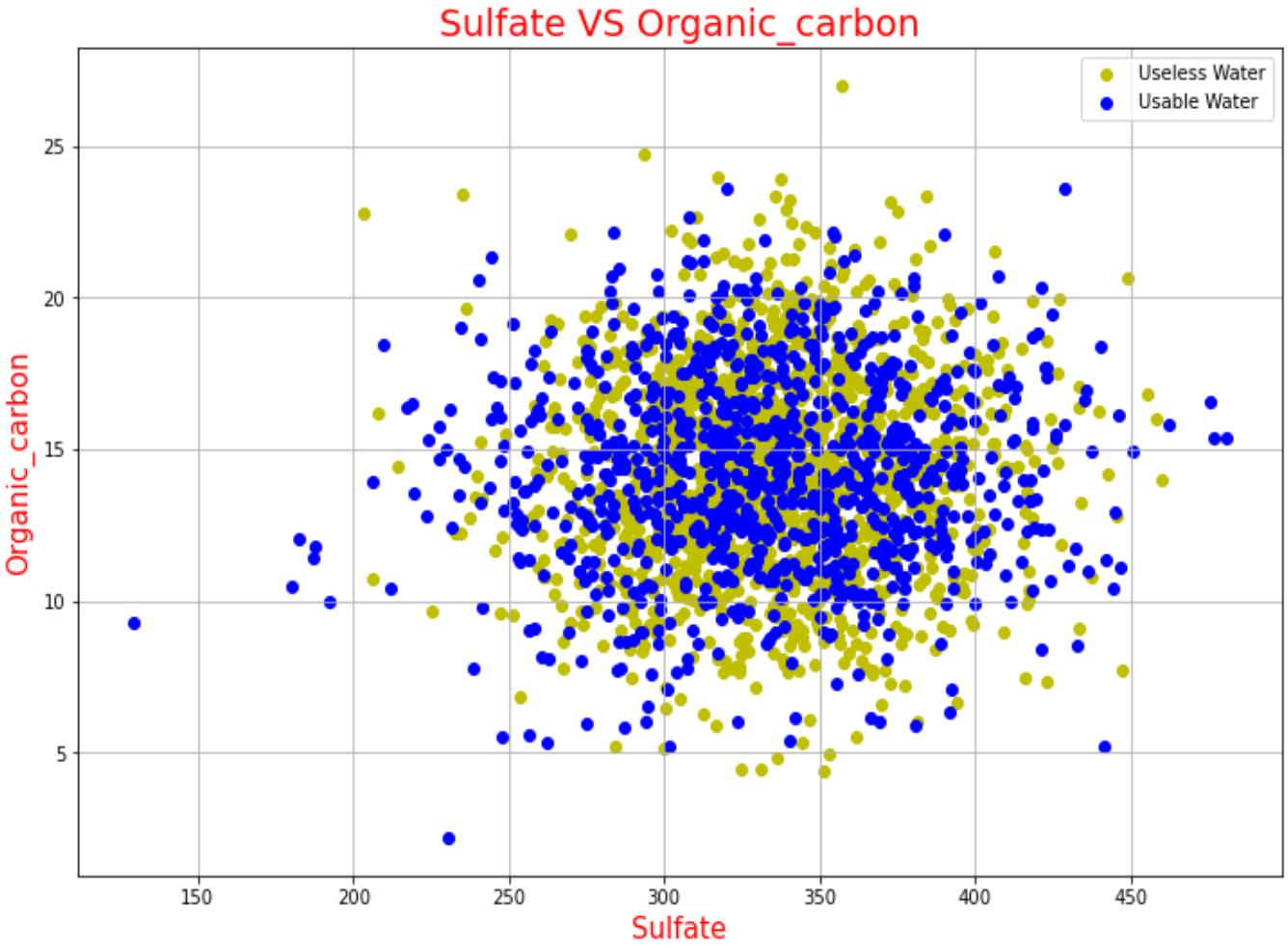



شکل بالا میزان ph آب را در مقایسه با مقدار تری هالومتان ها در آب بر حسب میکروگرم در لیتر نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در میزان ph آب بین ۴ تا ۱۰ و مقدار تری هالومتان بین ۵۰ تا ۹۰ بیشینه است. همچنین چگالی دیتاهای آب قابل استفاده در مقدار تری هالومتان بین ۵۰ تا ۹۰ و مقدار ph بین ۴ تا ۱۰ نسبت به چگالی دیتاهای آب غیرقابل استفاده در همین بازه ها بیشتر است.

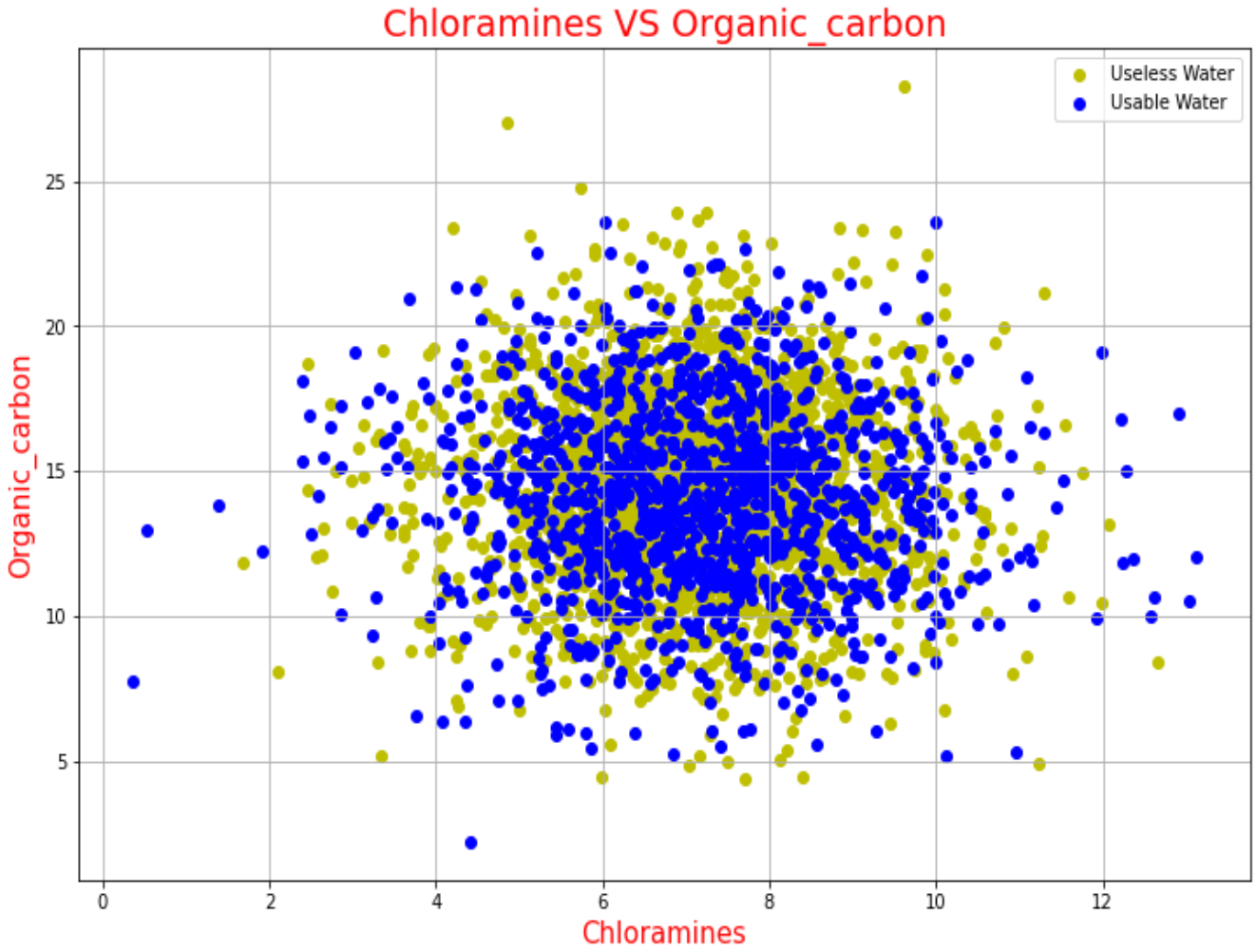



شکل بالا میزان هدایت الکتریکی آب را در مقایسه با مقدار کربن آلی در آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند.

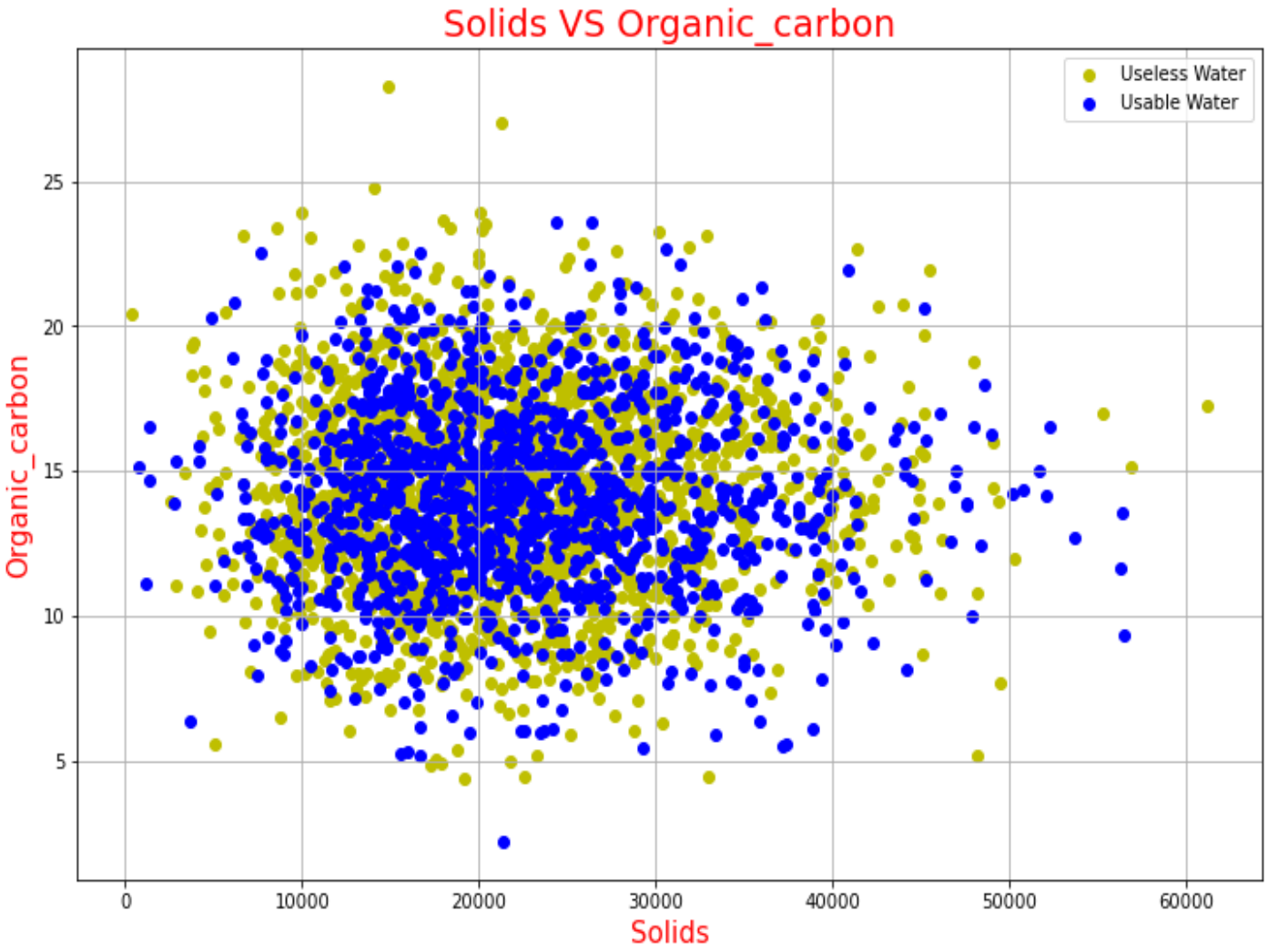
چگالی دیتاها در میزان هدایت الکتریکی آب بین ۳۰۰ تا ۵۰۰ و مقدار کربن آلی بین ۱۰ تا ۲۰ بیشینه است.



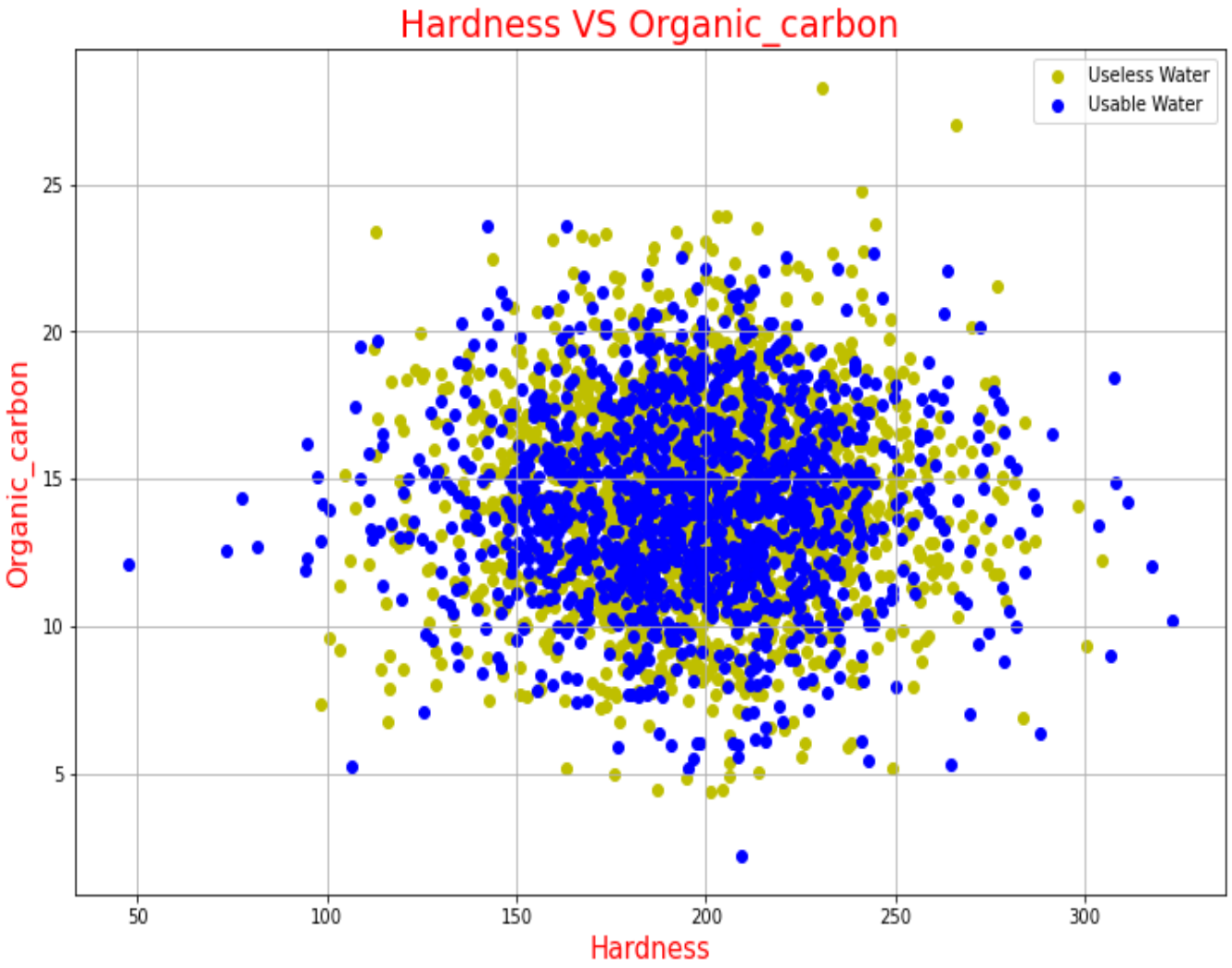

شکل بالا میزان سولفات های محلول در آب را  در مقایسه با مقدار کربن آلی در آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در میزان سولفات آب بین ۳۰۰ تا ۴۰۰ و مقدار کربن آلی بین ۱۰ تا ۲۰ بیشینه است.



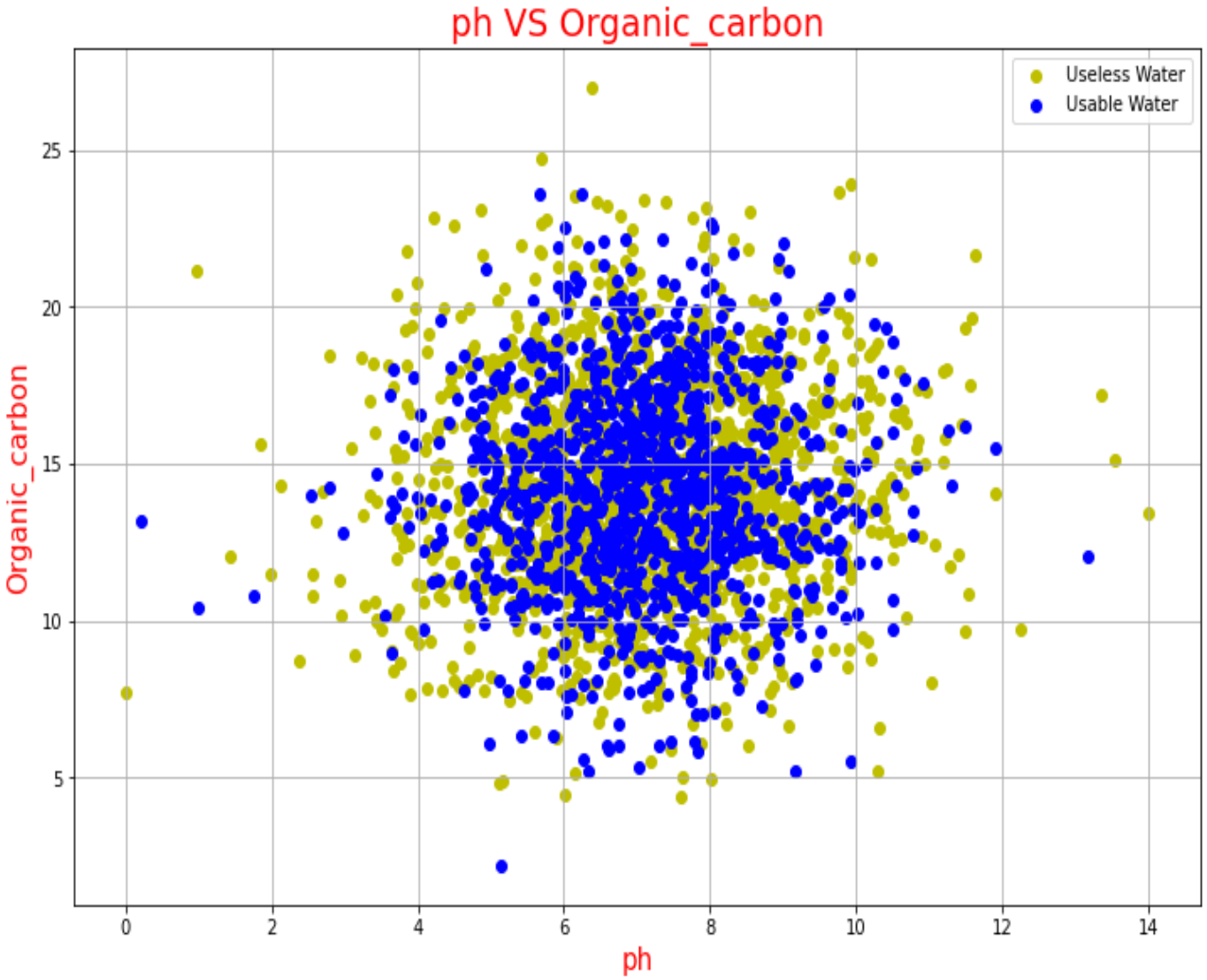

شکل بالا میزان کلرامین موجود در آب را  در مقایسه با مقدار کربن آلی در آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند.
چگالی دیتاها در میزان کلرامین آب بین ۵ تا ۹ و مقدار کربن آلی بین ۱۰ تا ۲۰ بیشینه است.



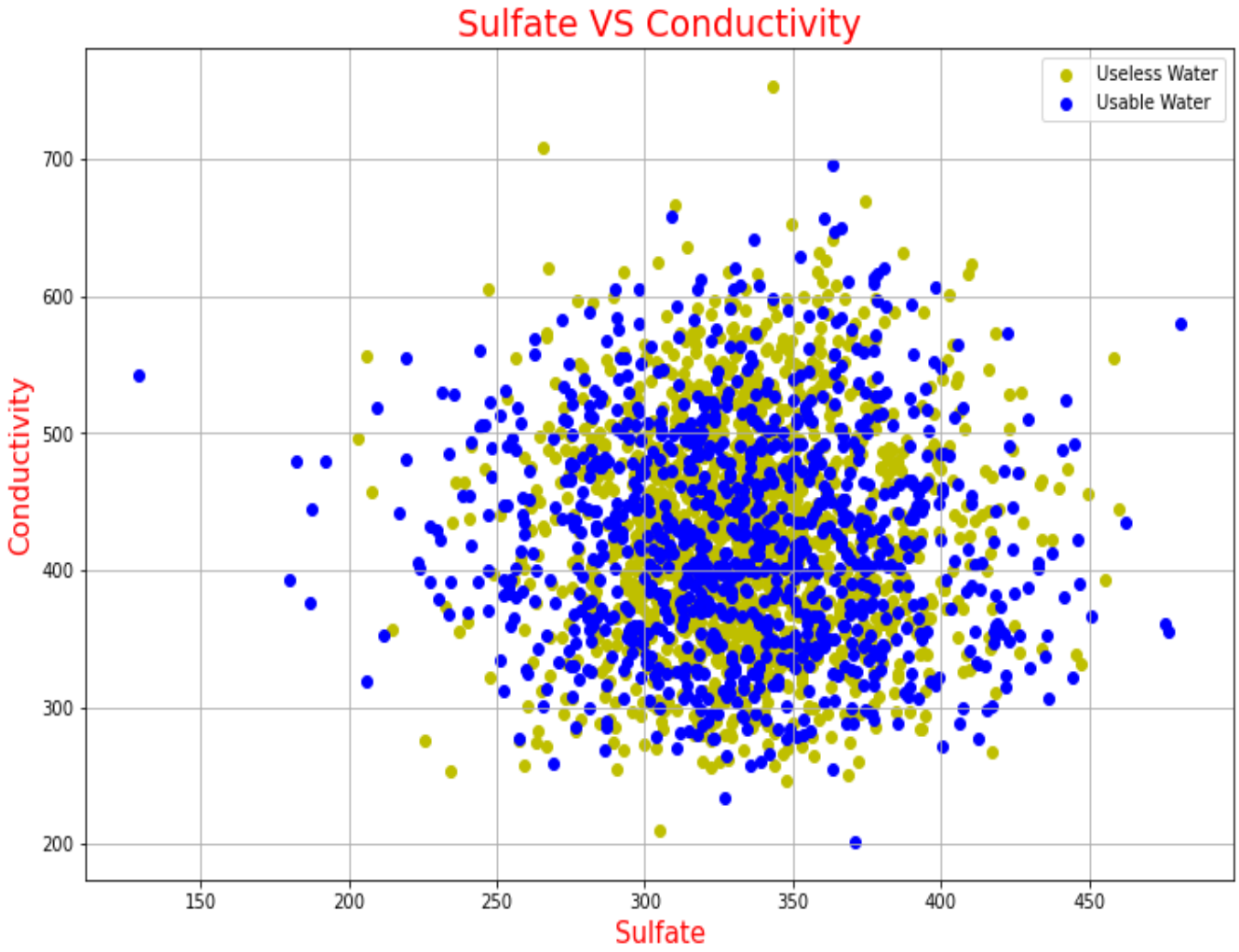

شکل بالا میزان کل مواد جامد موجود در آب را  در مقایسه با مقدار کربن آلی در آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در میزان مواد جامد آب بین ۱۰۰۰۰ تا ۳۰۰۰۰ و مقدار کربن آلی بین ۱۰ تا ۲۰ بیشینه است.



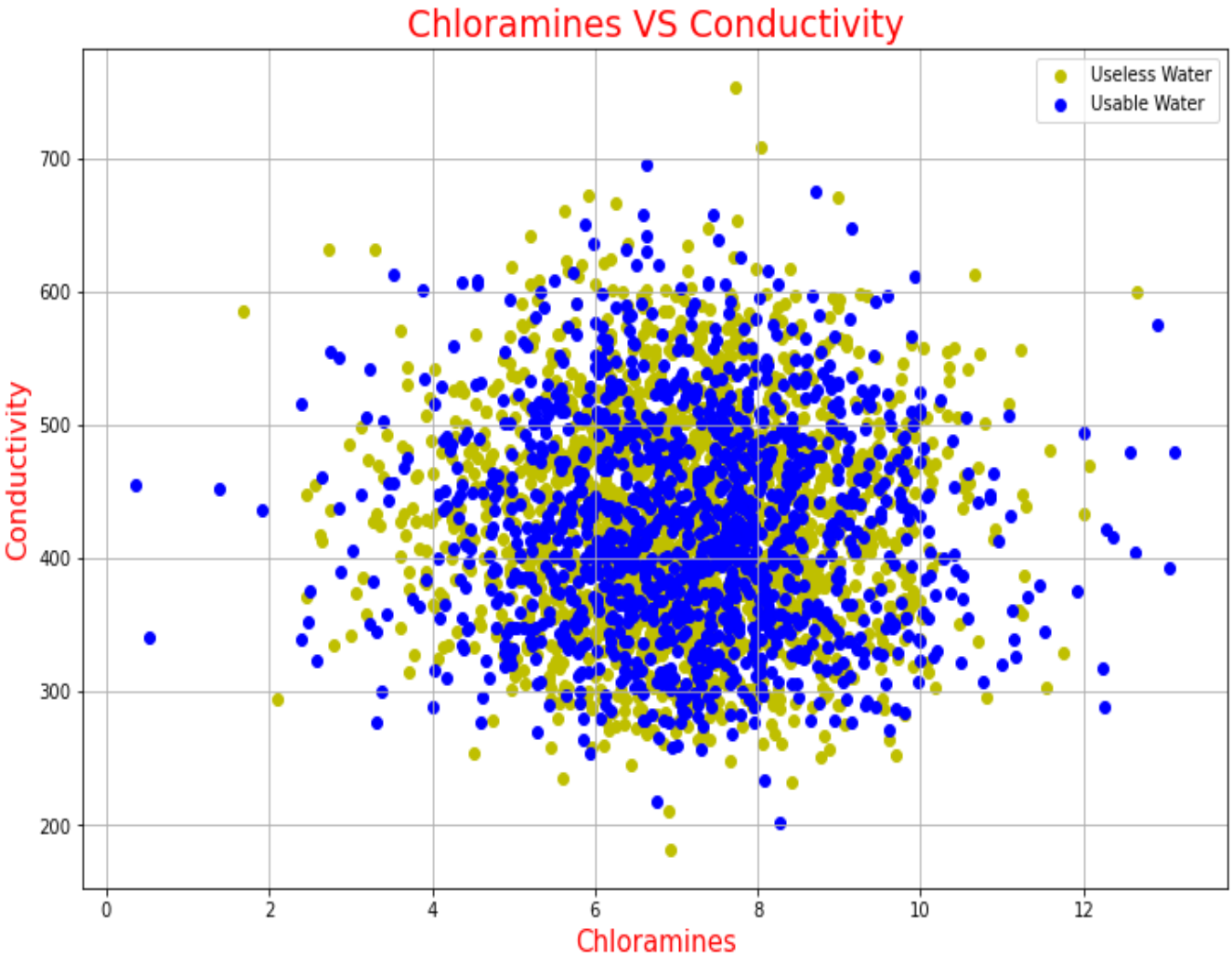

شکل بالا میزان سختی آب را  در مقایسه با مقدار کربن آلی در آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در میزان سختی آب بین ۱۵۰ تا ۲۵۰ و مقدار کربن آلی بین ۱۰ تا ۲۰ بیشینه است.



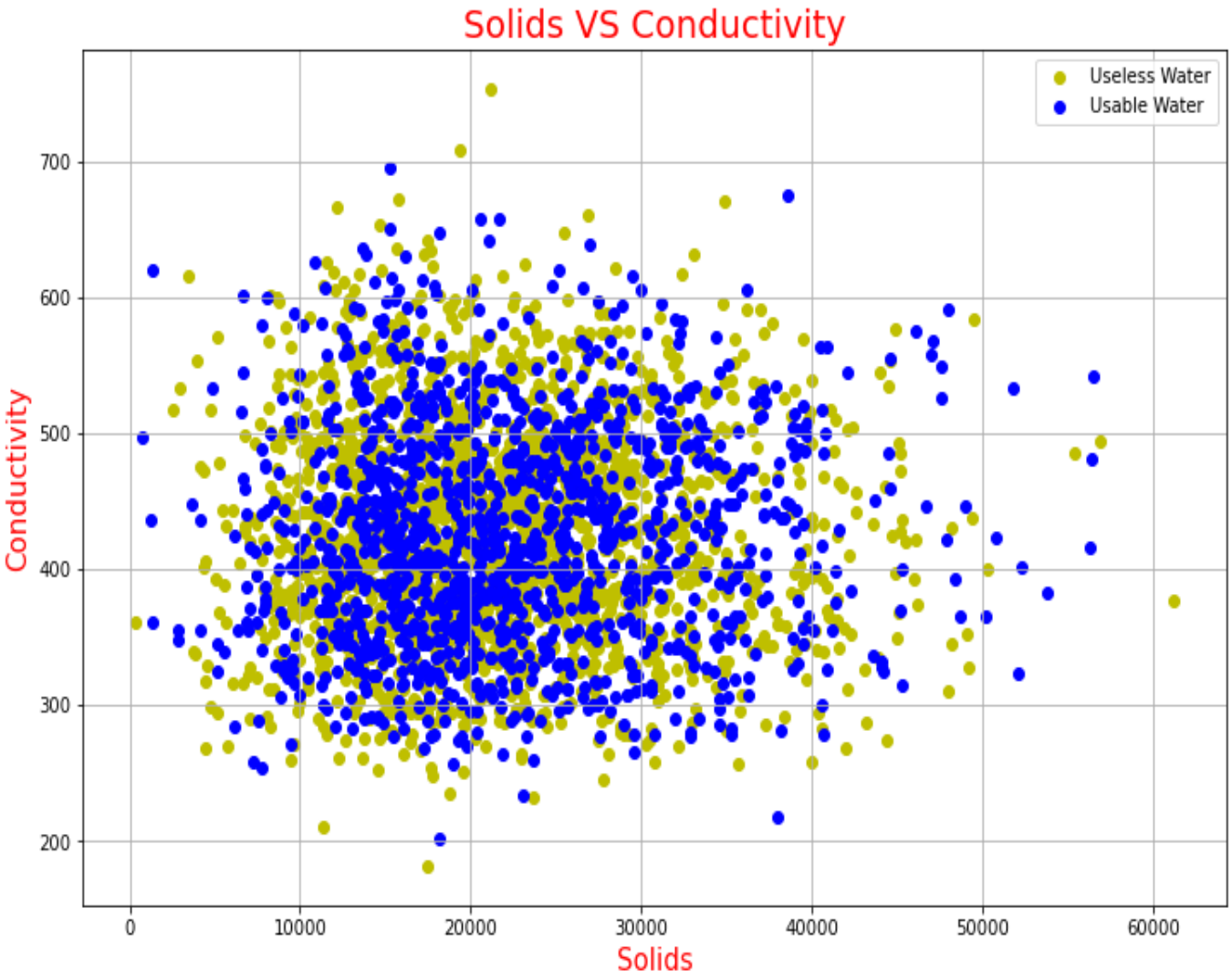

شکل بالا میزان ph آب را در مقایسه با مقدار کربن آلی در آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در میزان ph آب بین ۴ تا ۱۰ و مقدار کربن آلی بین ۱۰ تا ۲۰ بیشینه است. همچنین با پیشروی در نمودار به سمت ph های بیشتر یا کمتر از این بازه به تعداد دیتاهای آب غیرقابل استفاده افزوده و از تعداد دیتاهای آب قابل استفاده کاسته میشود.



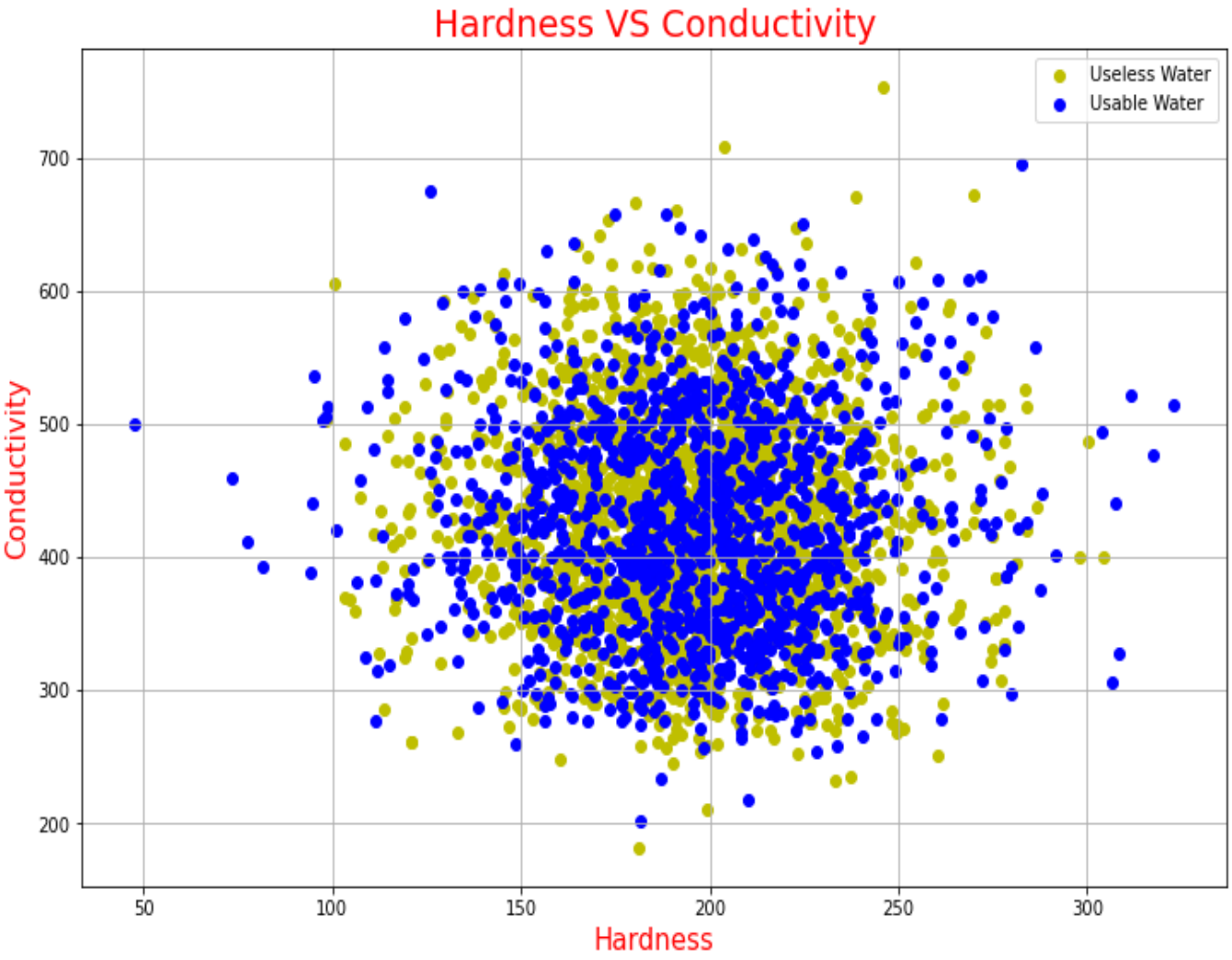

شکل بالا مقدار سولفات های محلول در آب را  در مقایسه با میزان هدایت الکتریکی آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در مقدار سولفات بین ۳۰۰ تا ۴۰۰ و میزان هدایت الکتریکی بین ۳۰۰ تا ۵۵۰ بیشینه است.



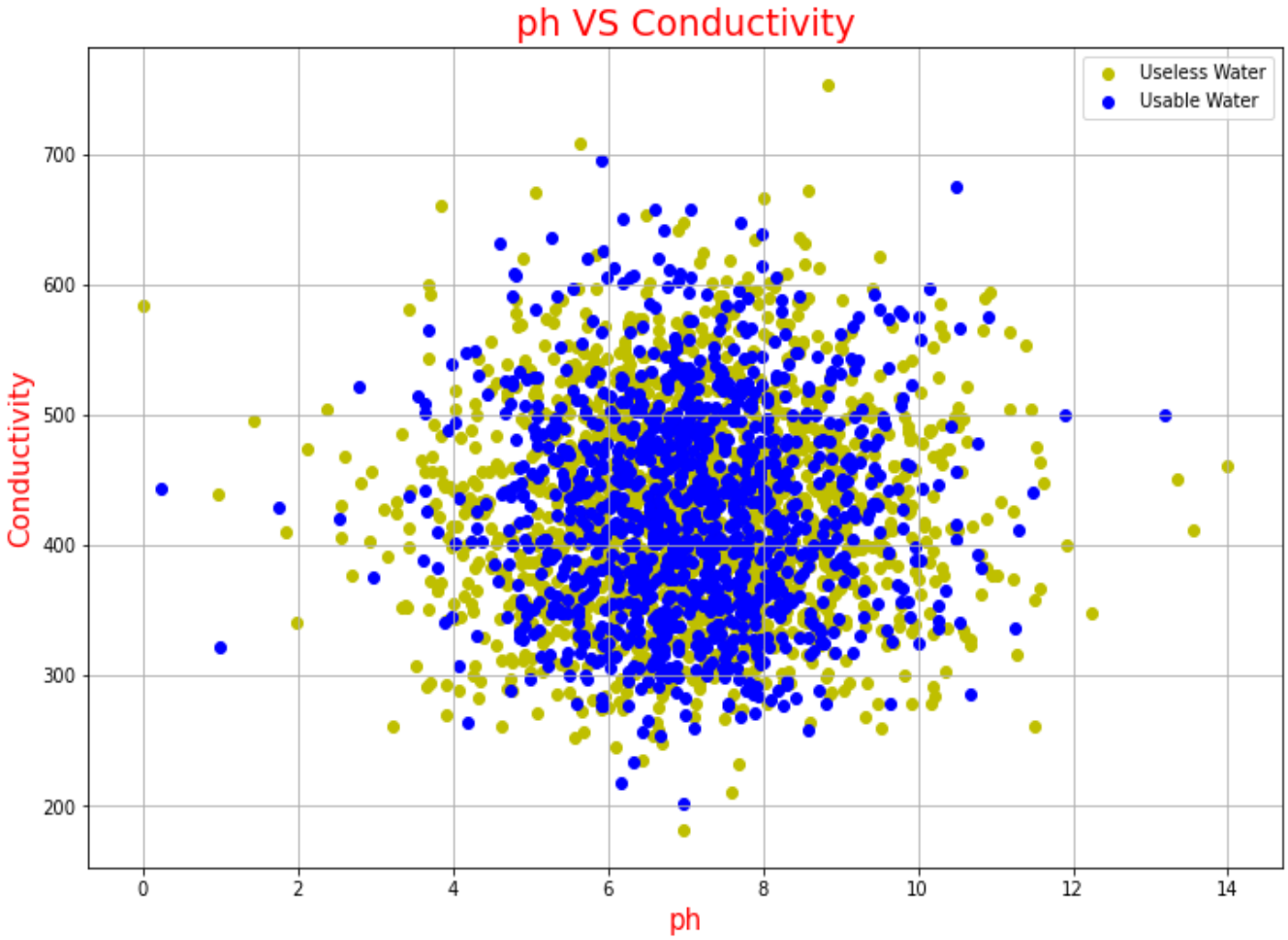

شکل بالا مقدار کلرامین موجود در آب را  در مقایسه با میزان هدایت الکتریکی آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در مقدار کلرامین بین ۵ تا ۹ و میزان هدایت الکتریکی بین ۳۰۰ تا ۵۵۰ بیشینه است.



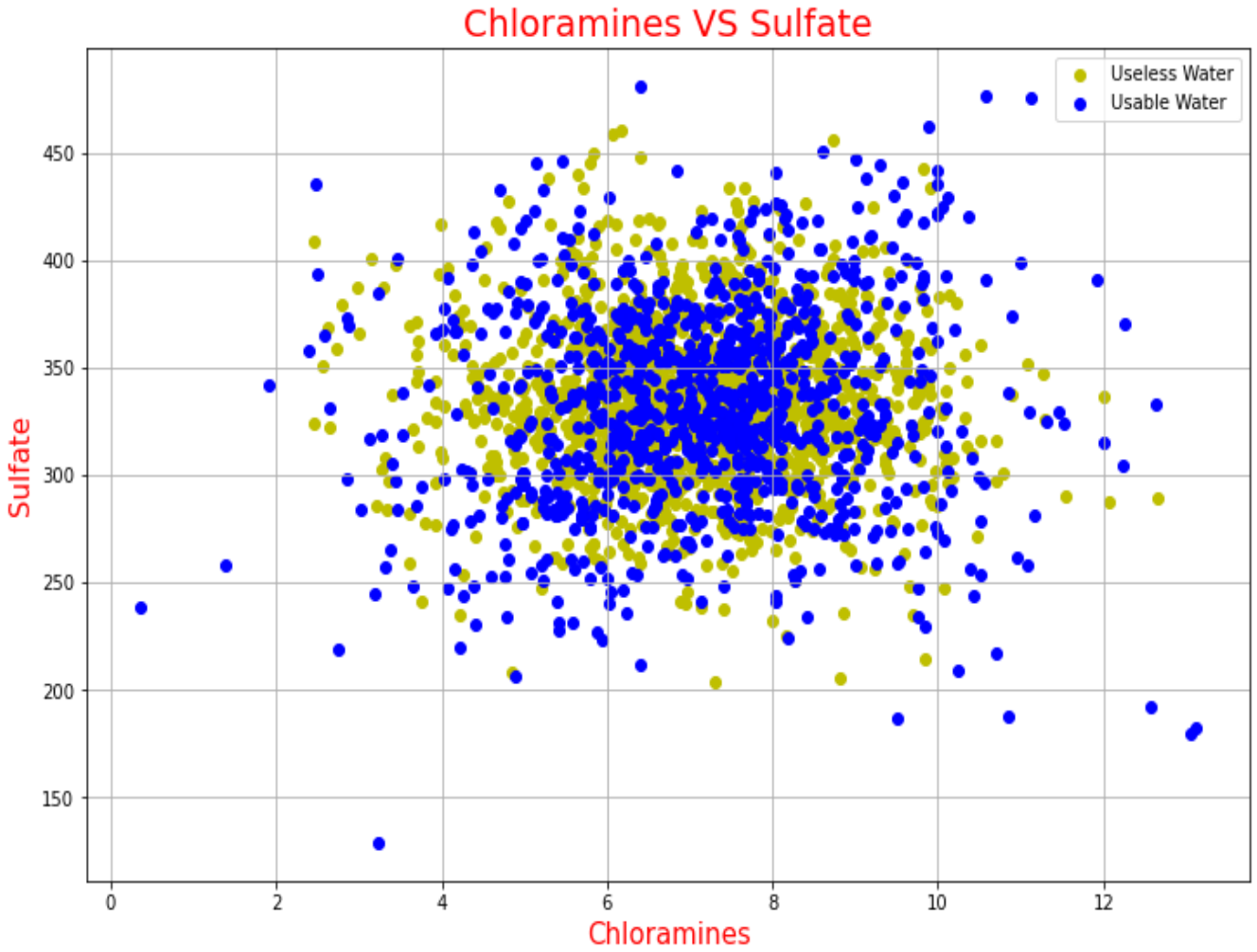

شکل بالا مقدار کل مواد جامد موجود در آب را  در مقایسه با میزان هدایت الکتریکی آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در مقدار مواد جامد بین ۱۰۰۰۰ تا ۳۰۰۰۰ و میزان هدایت الکتریکی بین ۳۰۰ تا ۵۵۰ بیشینه است.



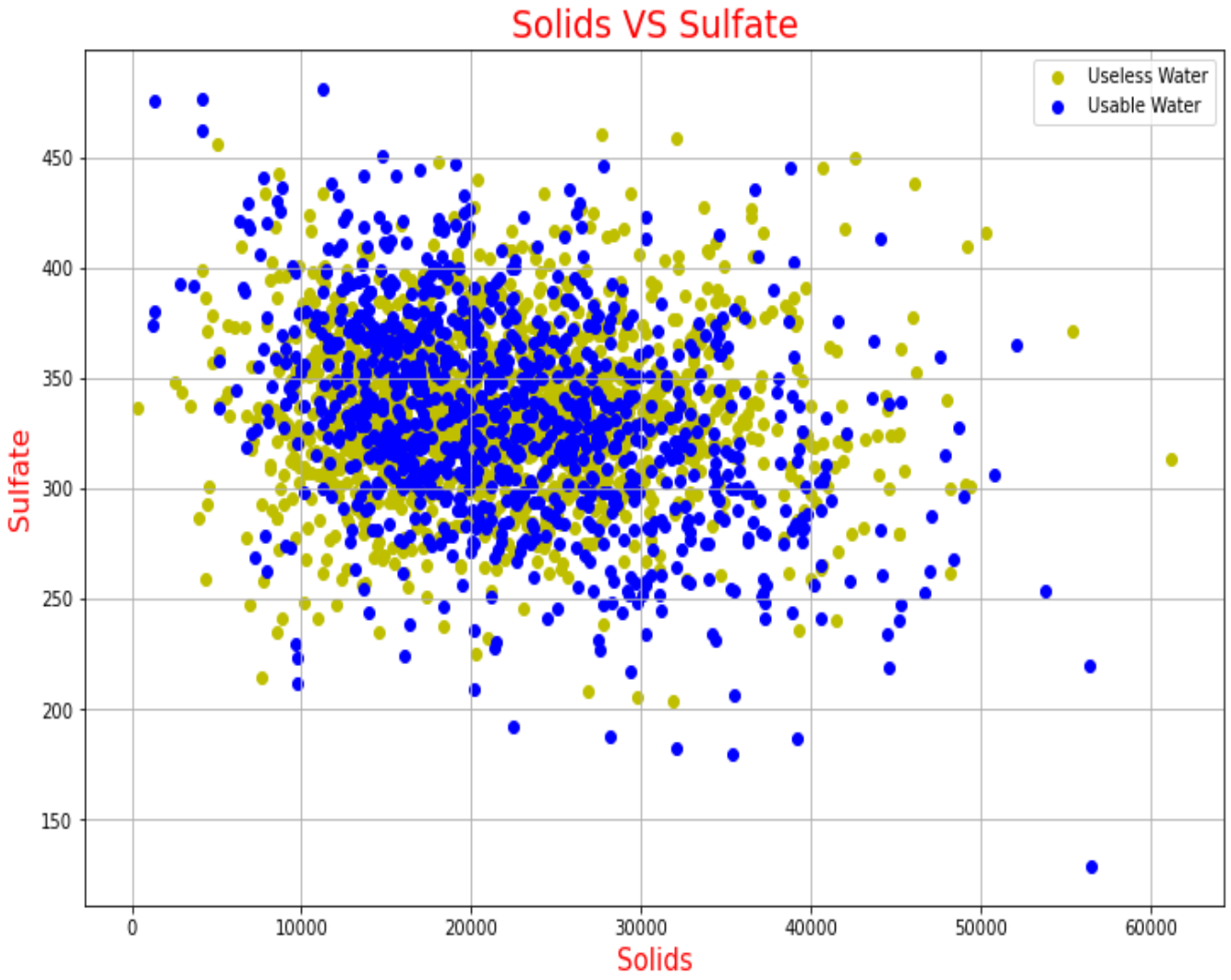

شکل بالا میزان سختی آب را  در مقایسه با میزان هدایت الکتریکی آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در میزان سختی بین ۱۵۰ تا ۲۵۰ و میزان هدایت الکتریکی بین ۳۰۰ تا ۵۵۰ بیشینه است.



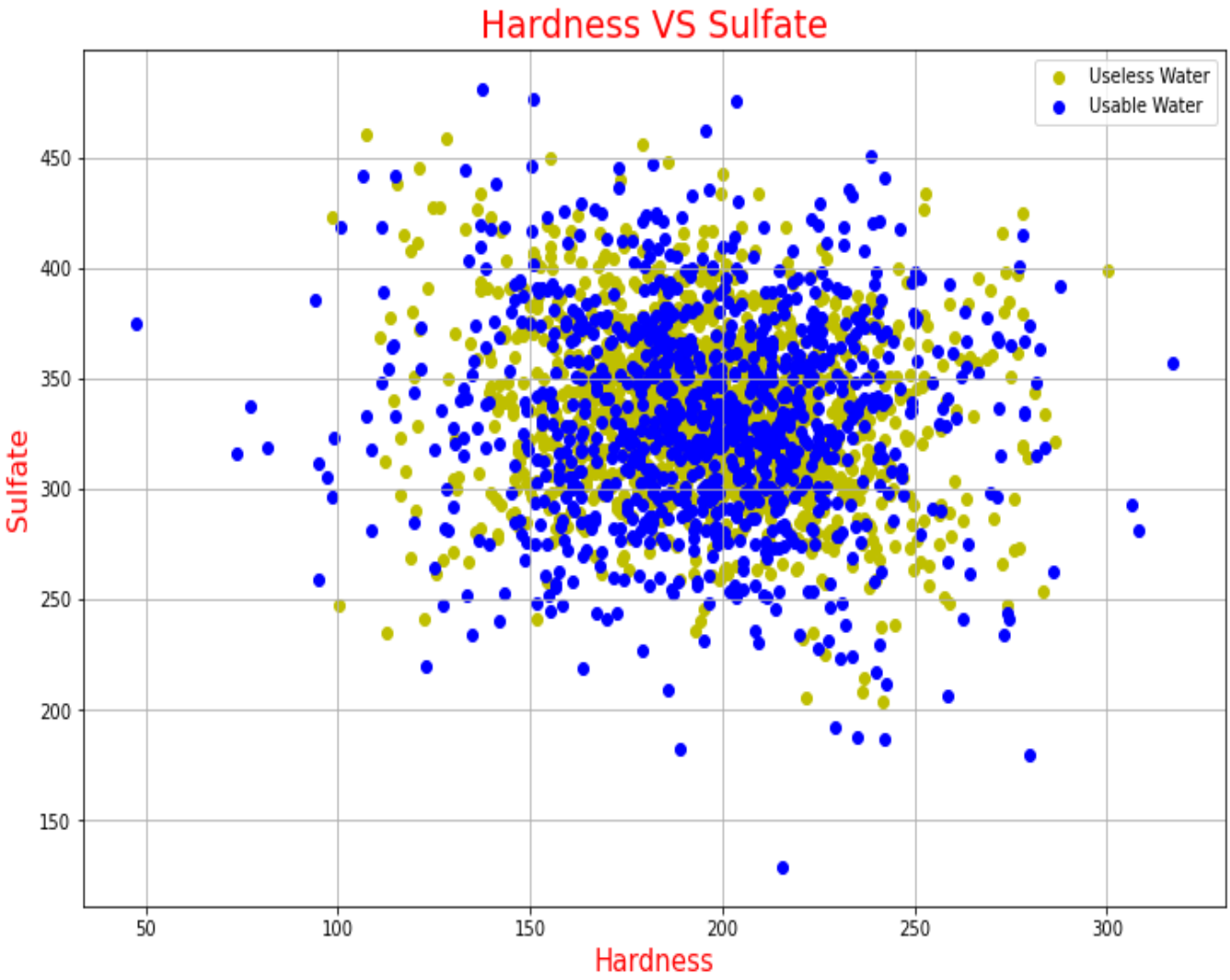

شکل بالا میزان ph آب را در مقایسه با میزان هدایت الکتریکی آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در میزان ph بین ۴ تا ۱۰ و میزان هدایت الکتریکی بین ۳۰۰ تا ۵۵۰ بیشینه است. همچنین با پیشروی در نمودار به سمت ph های بیشتر یا کمتر از این بازه به تعداد دیتاهای آب غیرقابل استفاده افزوده و از تعداد دیتاهای آب قابل استفاده کاسته میشود.



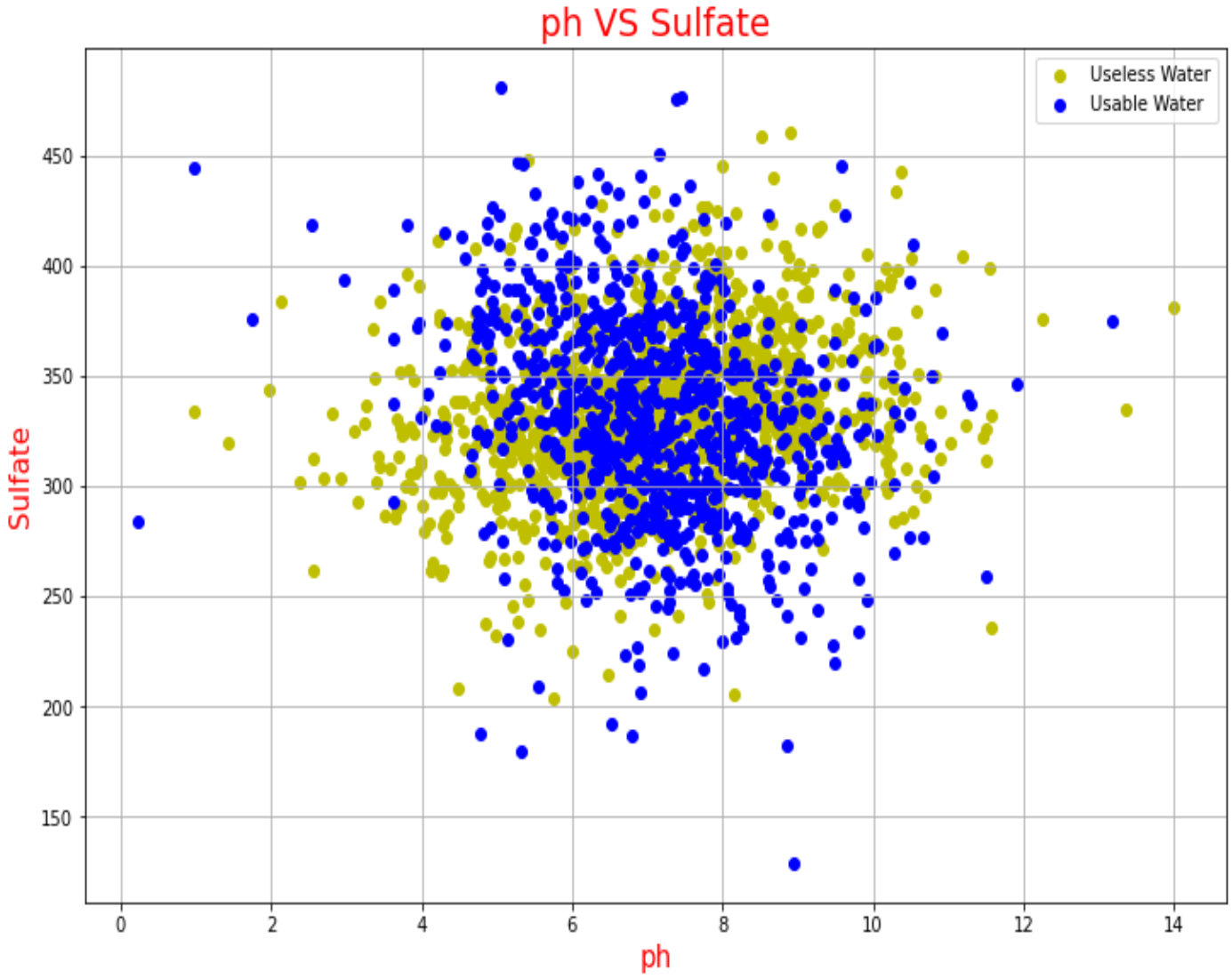

شکل بالا مقدار کلرامین موجود در آب را  در مقایسه با مقدار سولفات های محلول در آب برحسب میلی گرم در لیتر نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در مقدار کلرامین بین ۵ تا ۹ و مقدار سولفات بین ۳۰۰ تا ۴۰۰ بیشینه است.




شکل بالا مقدار کل مواد جامد موجود در آب را  در مقایسه با مقدار سولفات های محلول در آب برحسب میلی گرم در لیتر نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در مقدار مواد جامد بین ۱۰۰۰۰ تا ۳۰۰۰۰ و مقدار سولفات بین ۳۰۰ تا ۴۰۰ بیشینه است.

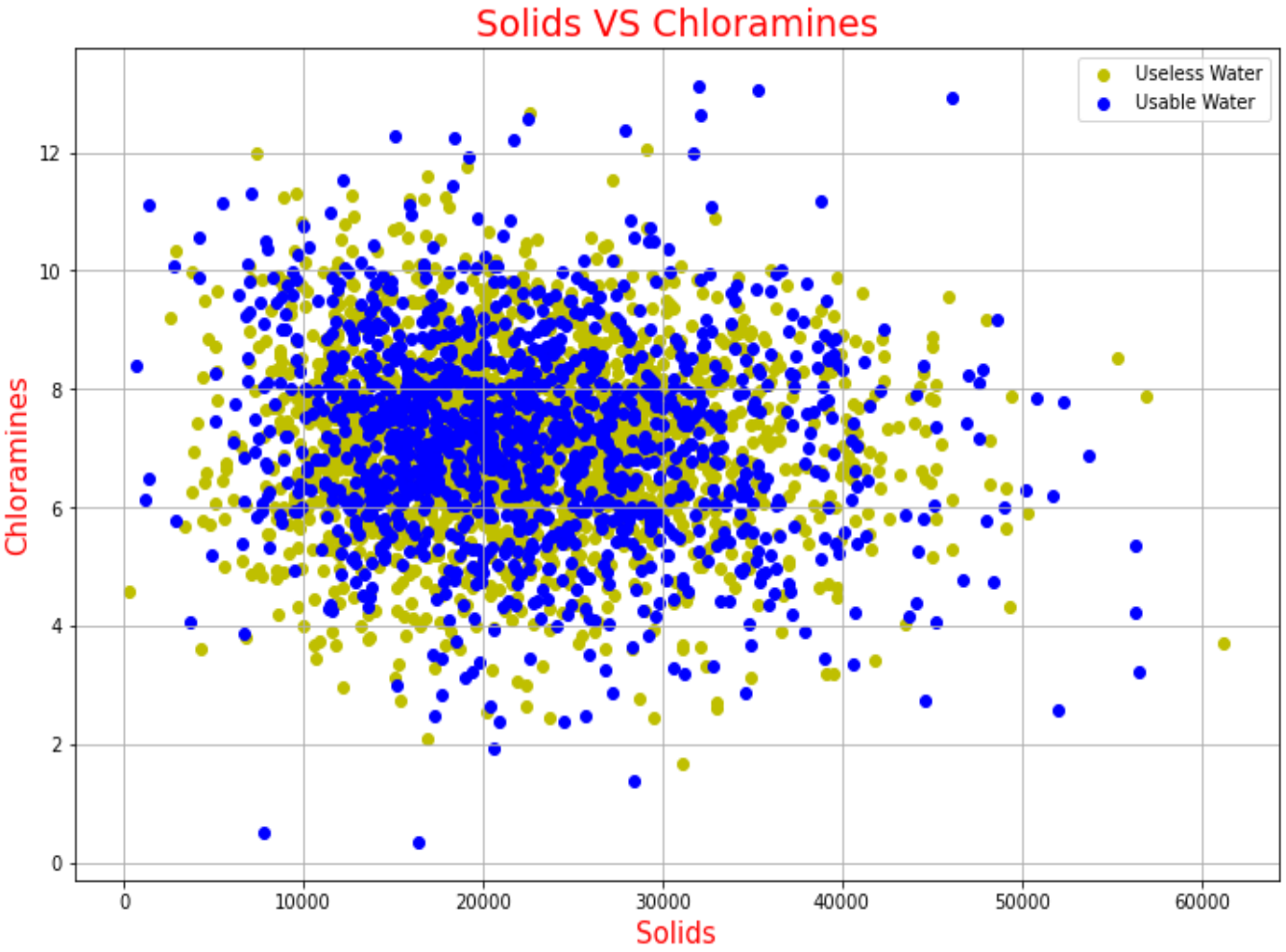



شکل بالا مقدار سختی آب را  در مقایسه با مقدار سولفات های محلول در آب برحسب میلی گرم در لیتر نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در میزان سختی بین ۱۵۰ تا ۲۵۰ و مقدار سولفات بین ۳۰۰ تا ۴۰۰ بیشینه است.

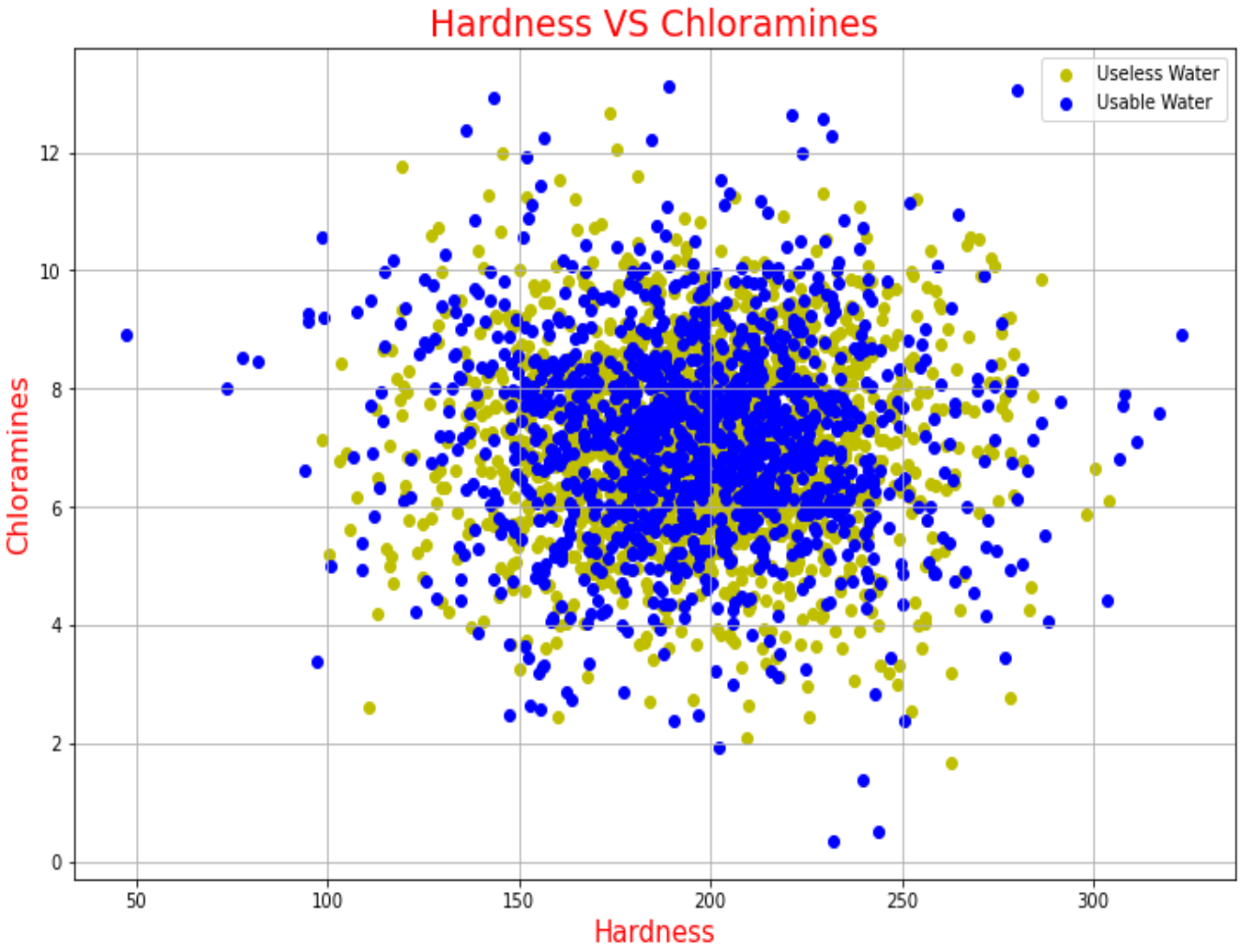



شکل بالا مقدار ph آب را  در مقایسه با مقدار سولفات های محلول در آب برحسب میلی گرم در لیتر نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در میزان ph بین ۴ تا ۱۰ و مقدار سولفات بین ۳۰۰ تا ۴۰۰ بیشینه است.

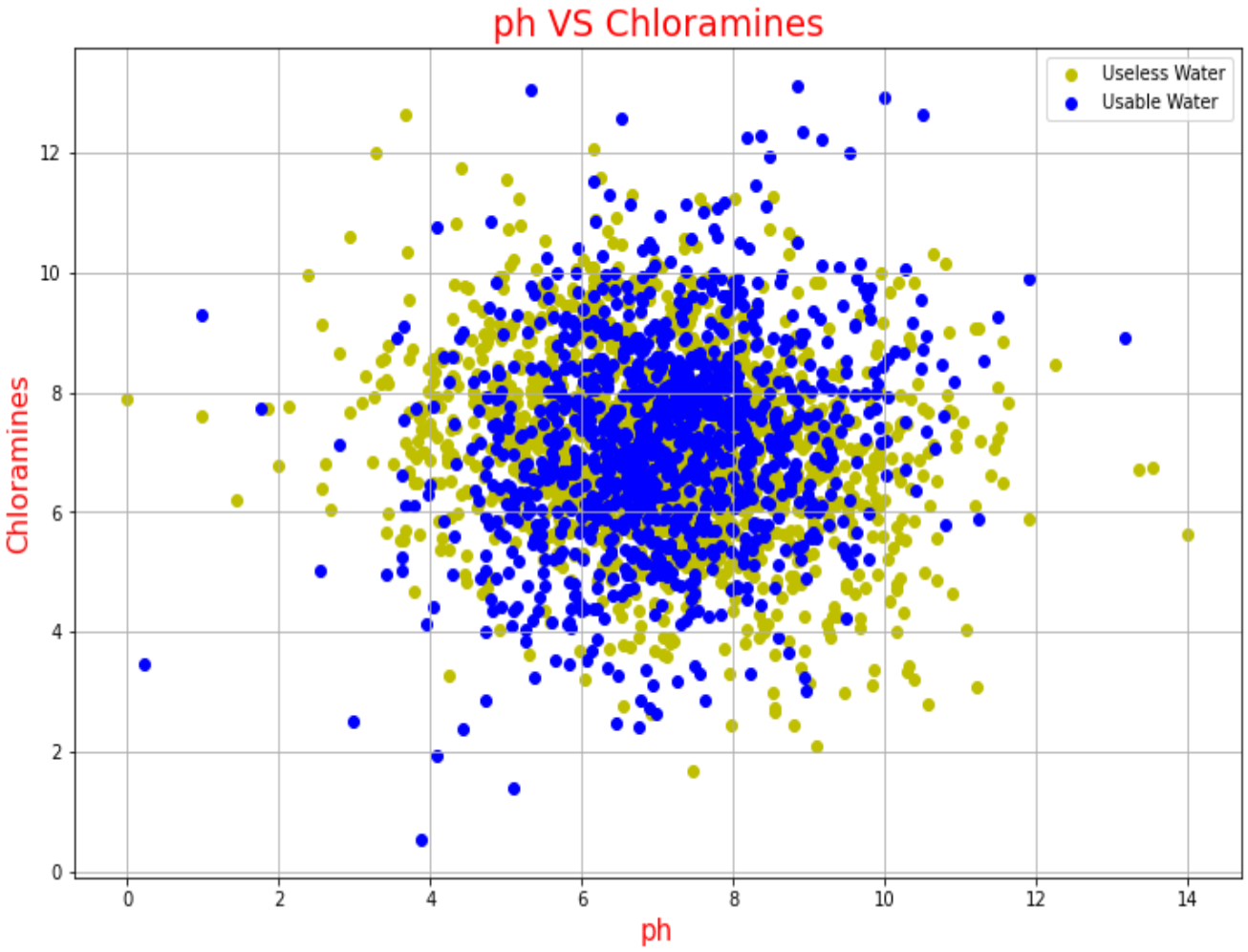
همچنین با پیشروی در نمودار به سمت  ph های بیشتر یا کمتر از این بازه به تعداد دیتاهای آب غیرقابل استفاده افزوده و از تعداد دیتاهای آب قابل استفاده کاسته میشود.



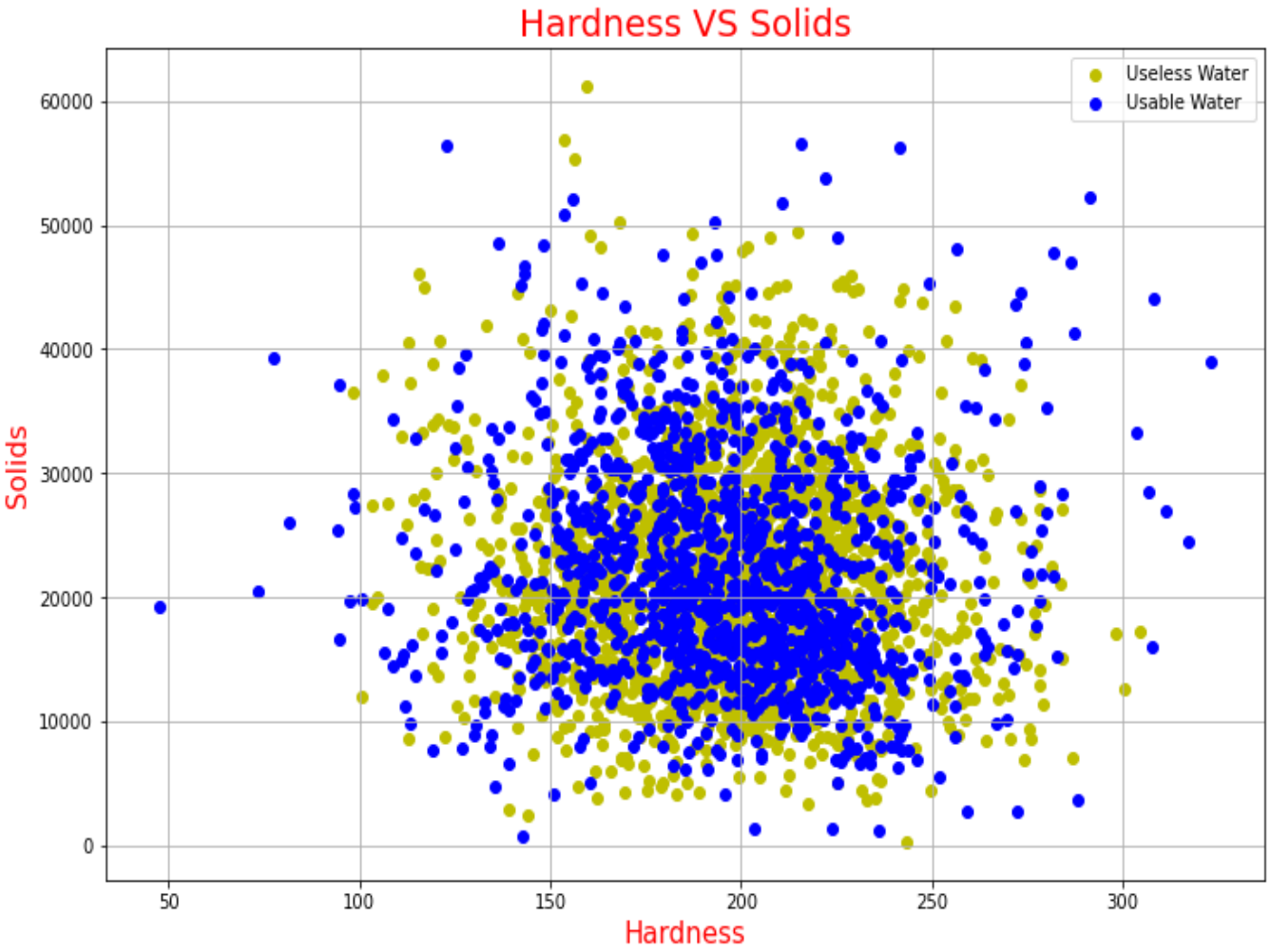

شکل بالا مقدار کل مواد جامد در آب را  در مقایسه با مقدار کلرامین موجود در آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در مقدار مواد جامد بین ۱۰۰۰۰ تا ۳۰۰۰۰ و مقدار کلرامین بین ۵ تا ۹ بیشینه است.



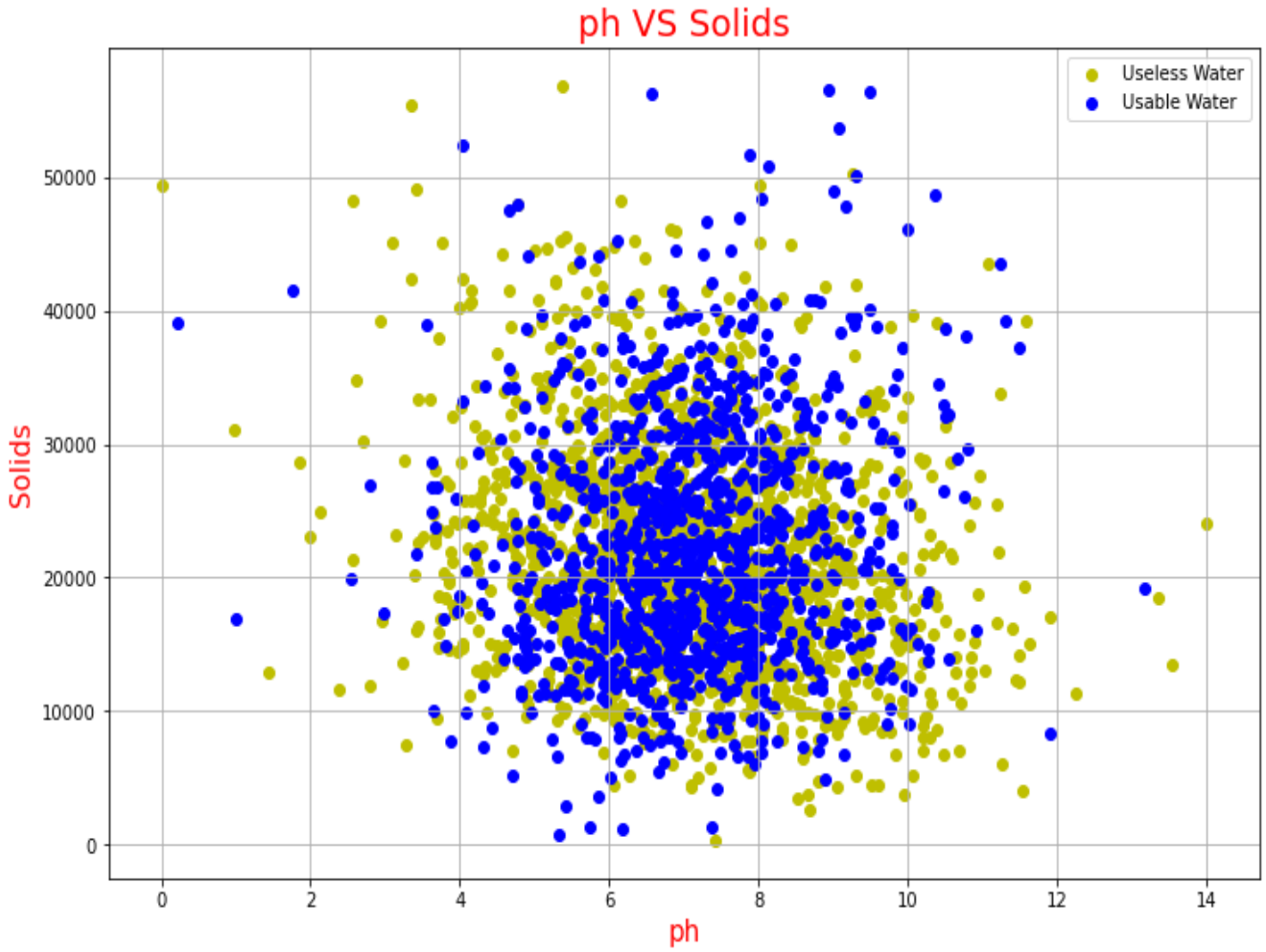

شکل بالا میزان سختی آب را  در مقایسه با مقدار کلرامین موجود در آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در میزان سختی بین ۱۵۰ تا ۲۵۰ و مقدار کلرامین بین ۵ تا ۹ بیشینه است.



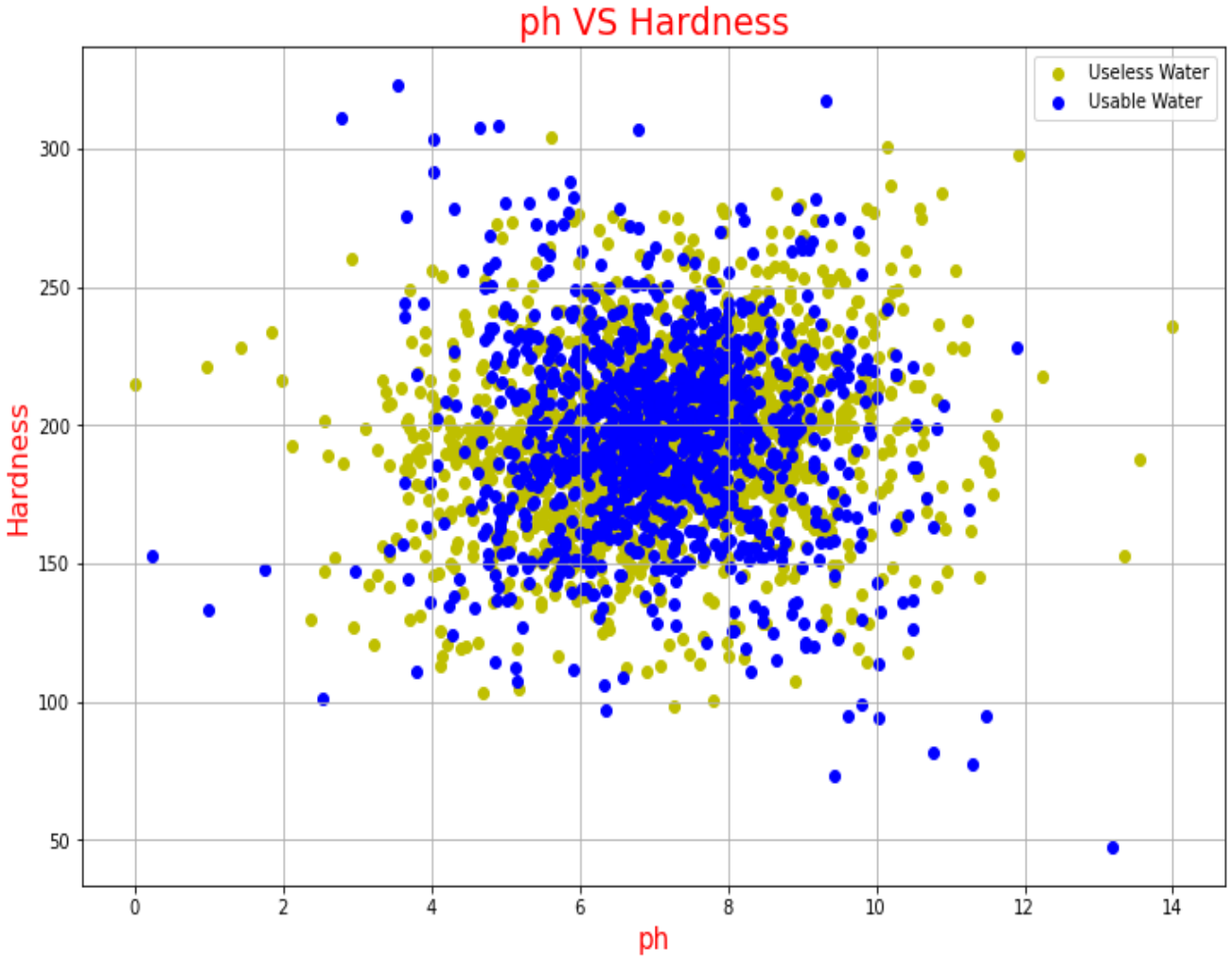

شکل بالا میزان ph آب را در مقایسه با مقدار کلرامین موجود در آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در میزان ph بین ۴ تا ۱۰ و مقدار کلرامین بین ۵ تا ۹ بیشینه است. همچنین با پیشروی در نمودار به سمت ph های بیشتر یا کمتر از این بازه به تعداد دیتاهای آب غیرقابل استفاده افزوده و از تعداد دیتاهای آب قابل استفاده کاسته میشود.




شکل بالا میزان سختی آب را  در مقایسه با مقدار کل مواد جامد موجود در آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند.
چگالی دیتاها در میزان سختی بین ۱۵۰ تا ۲۵۰ و مقدار مواد جامد بین ۱۰۰۰۰ تا ۳۰۰۰۰ بیشینه است.




شکل بالا میزان ph آب را در مقایسه با مقدار کل مواد جامد موجود در آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در میزان ph بین ۴ تا ۱۰ و مقدار مواد جامد بین ۱۰۰۰۰ تا ۳۰۰۰۰ بیشینه است. همچنین با پیشروی در نمودار به سمت ph های بیشتر یا کمتر از این بازه به تعداد دیتاهای آب غیرقابل استفاده افزوده و از تعداد دیتاهای آب قابل استفاده کاسته میشود.




شکل بالا میزان ph آب را در مقایسه با میزان سختی آب نشان می دهد، داده ها با کمک رنگ های زرد و آبی به دو گروه آب قابل استفاده و غیرقابل استفاده براساس تارگت مساله تقسیم شده اند. چگالی دیتاها در میزان ph بین ۴ تا ۱۰ و میزان سختی بین ۱۵۰ تا ۲۵۰ بیشینه است. همچنین با پیشروی در نمودار به سمت ph های بیشتر یا کمتر از این بازه به تعداد دیتاهای آب غیرقابل استفاده افزوده و از تعداد دیتاهای آب قابل استفاده کاسته میشود.

## مرحله دوم : بررسی خطا

### تشخیص نویزها و داده های پرت :

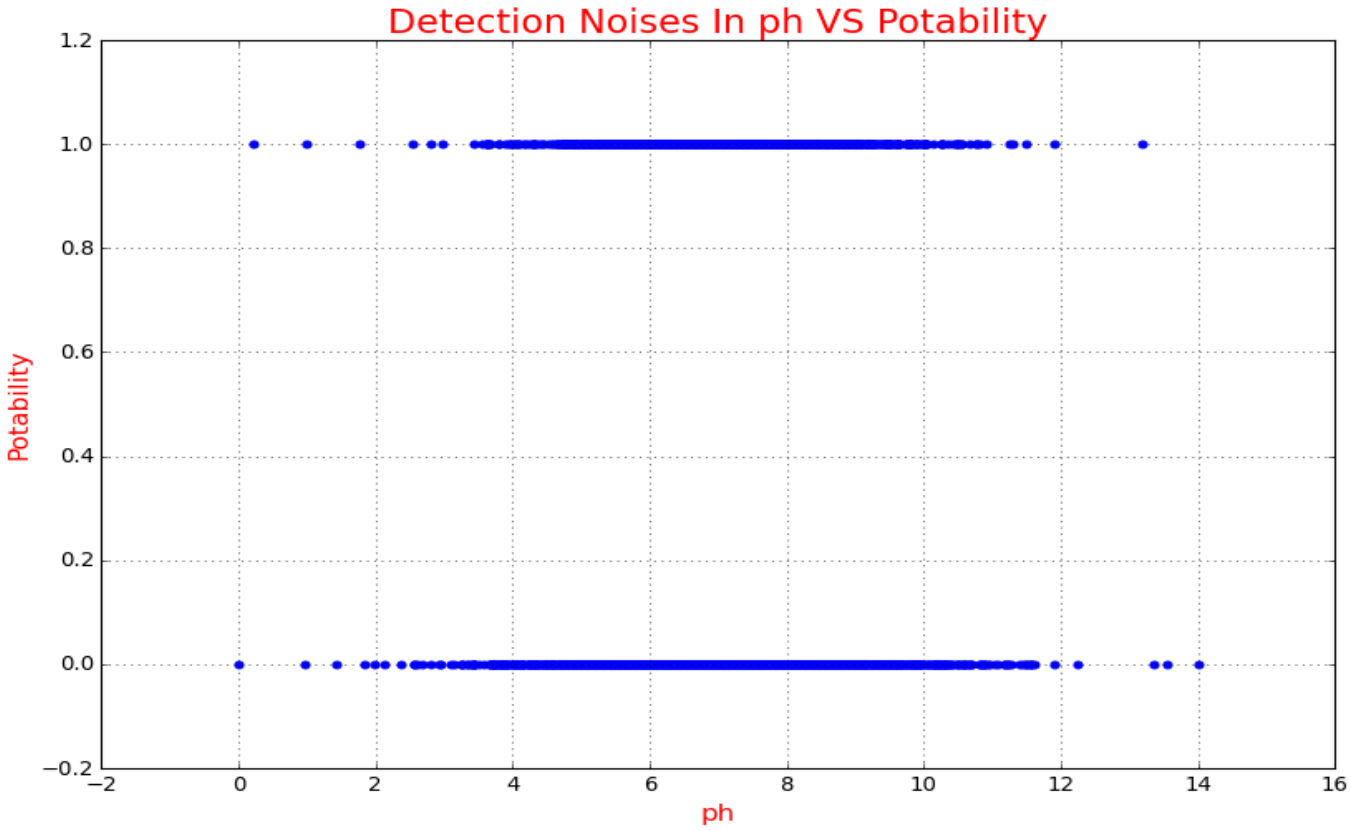
قبل از بررسی نویزها به علت وجود داده های گمشده در بعضی از ستون ها ابتدا با استفاده از دستور dropna داده های گمشده را از ستون مورد بررسی حذف می کنیم و سپس نمودار را ترسیم می کنیم.

در بررسی نویز بودن داده ها همیشه باید دو نکته را در نظر گرفت :

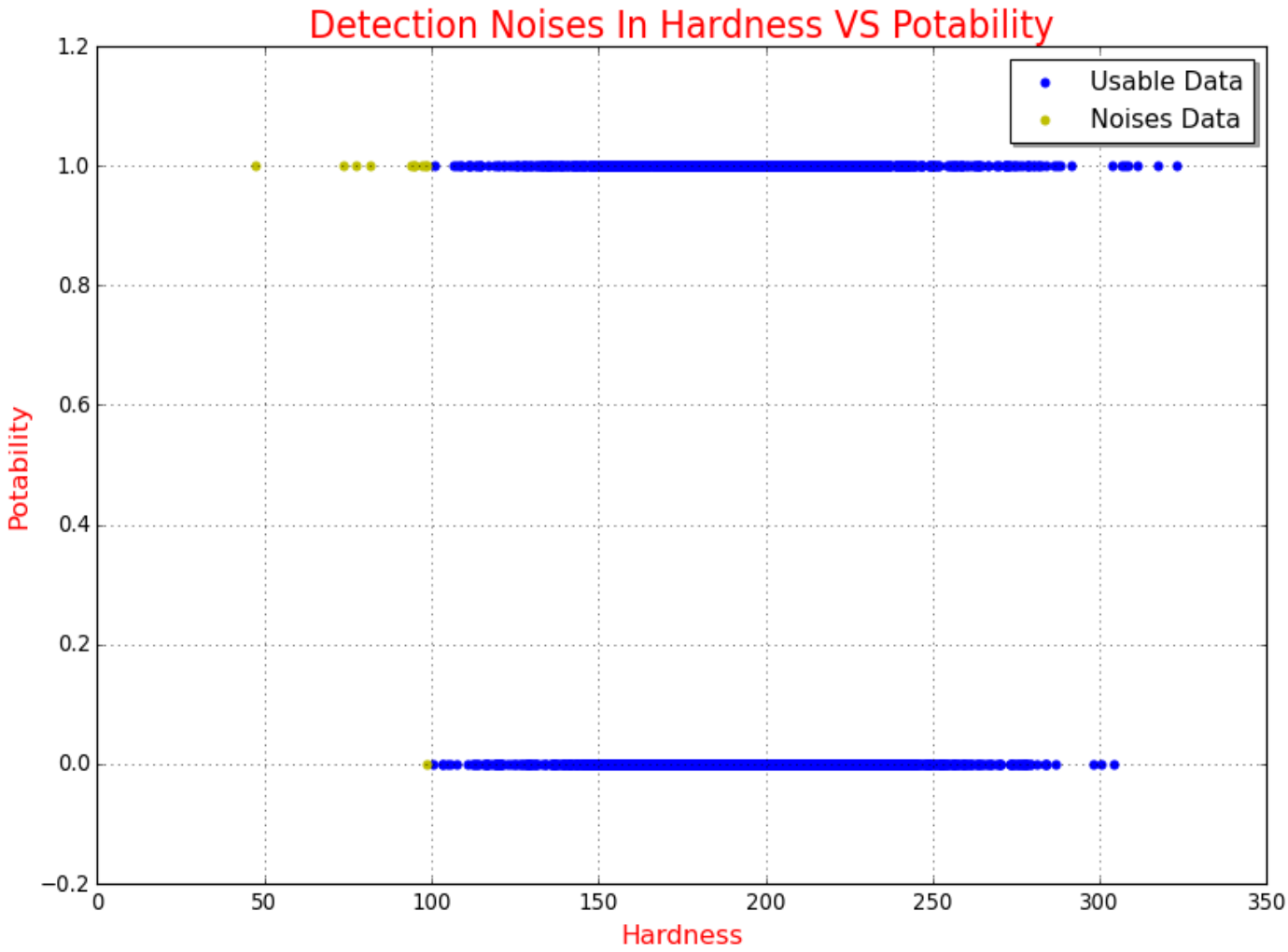
۱- میزان فاصله دیتا از تراکم دیتاها

۲- اعتبار داشتن از لحاظ علمی و منطقی

پس چنانچه داده ی پرتی با فاصله زیاد از سایر دیتاها پیدا شد نیاز به بررسی این دارد که آیا این فاصله از لحاظ علمی و منطقی قابل قبول است یا خیر.

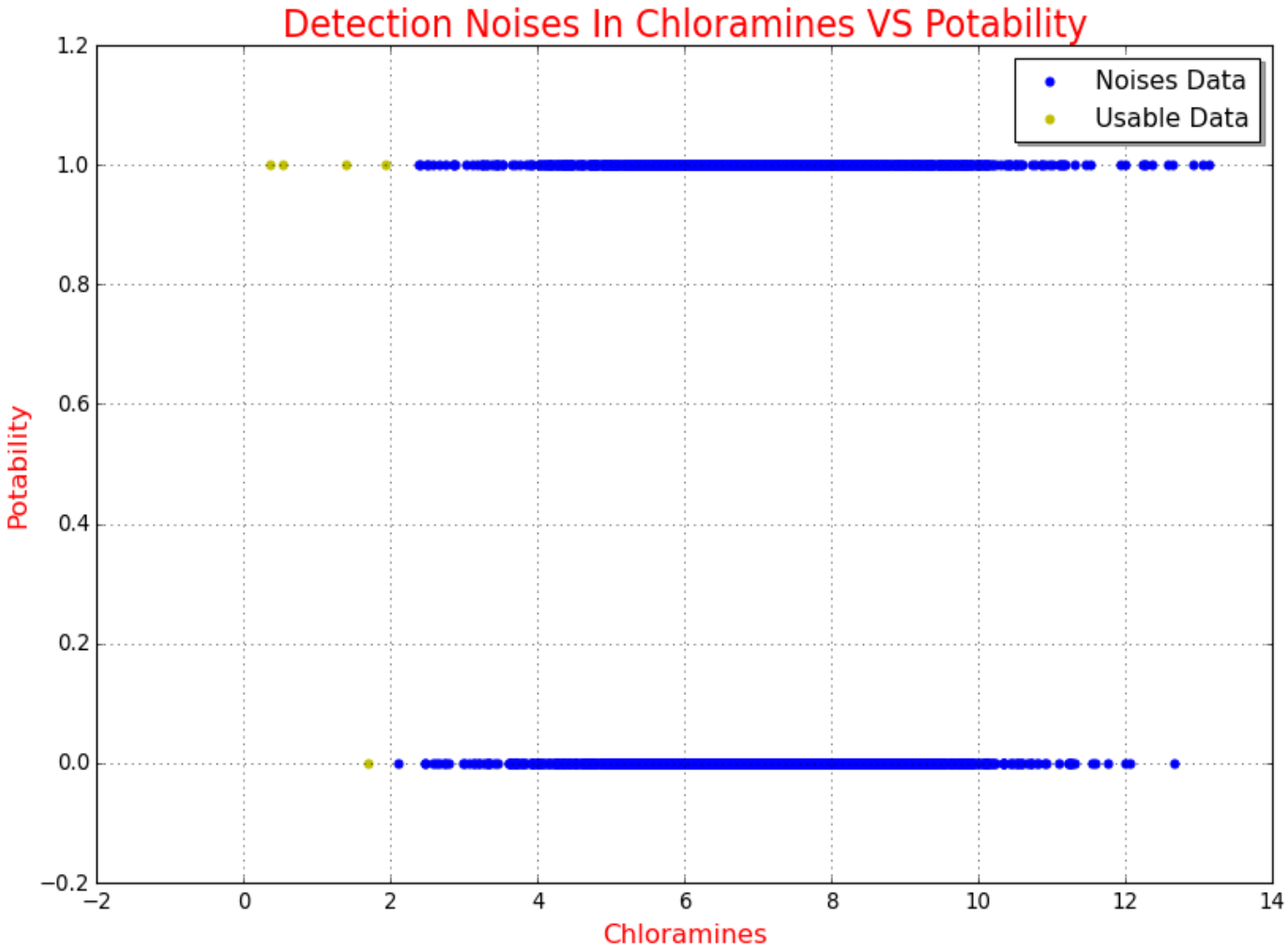



فیچر ph تعدادی داده ی پرت مشکوک به نویز دارد اما چون بازه قابل قبول برای ph بین ۰ تا ۱۴ است از لحاظ علمی و منطقی مورد تایید است.

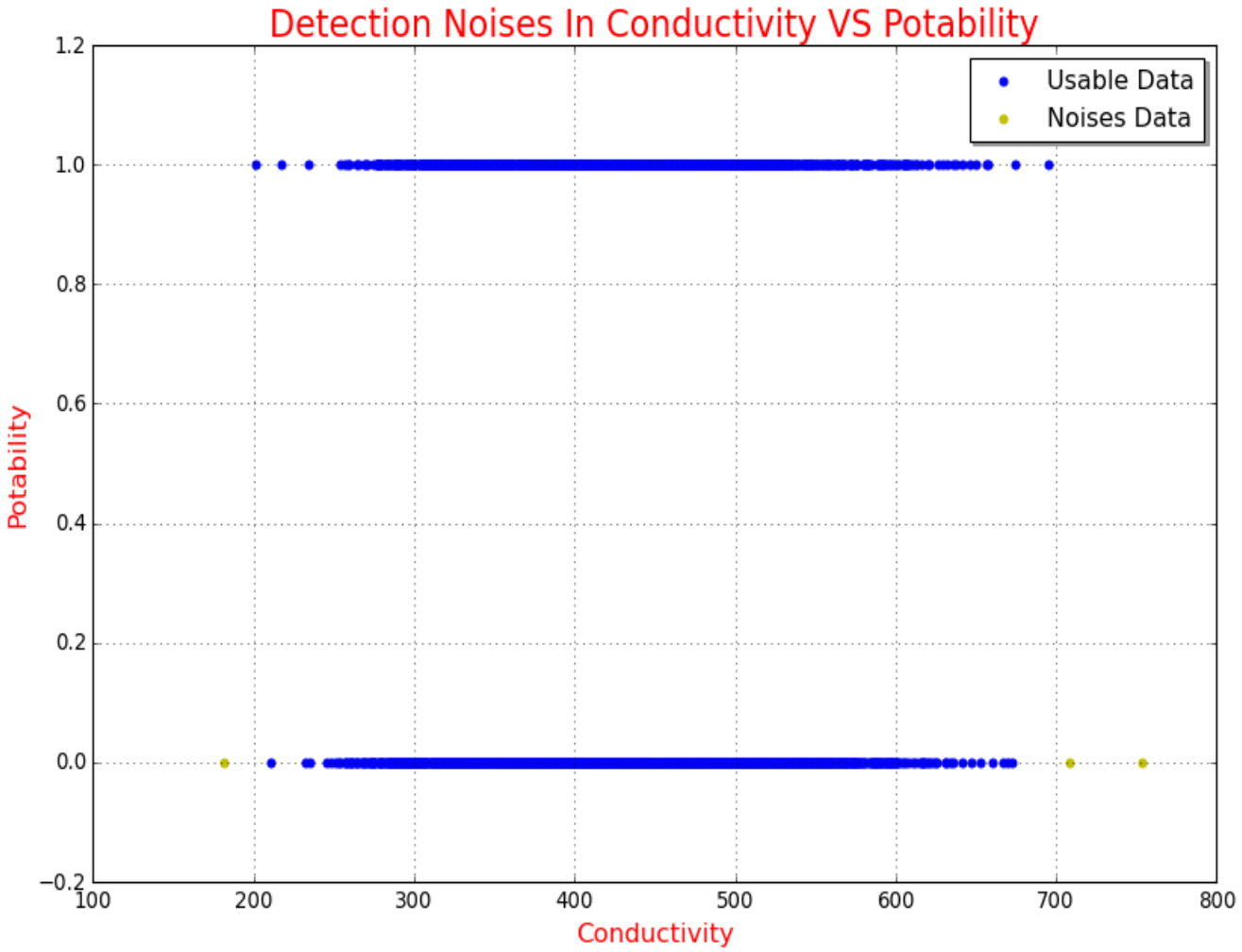



در شکل بالا دیتاهای نویز با رنگ زرد تفکیک شده اند با توجه به اینکه بازه سختی آب استخر بین ۱۰۰ تا ۶۰۰ قابل قبول است دیتاهای کمتر از ۱۰۰ نویز تشخیص داده شده اند.

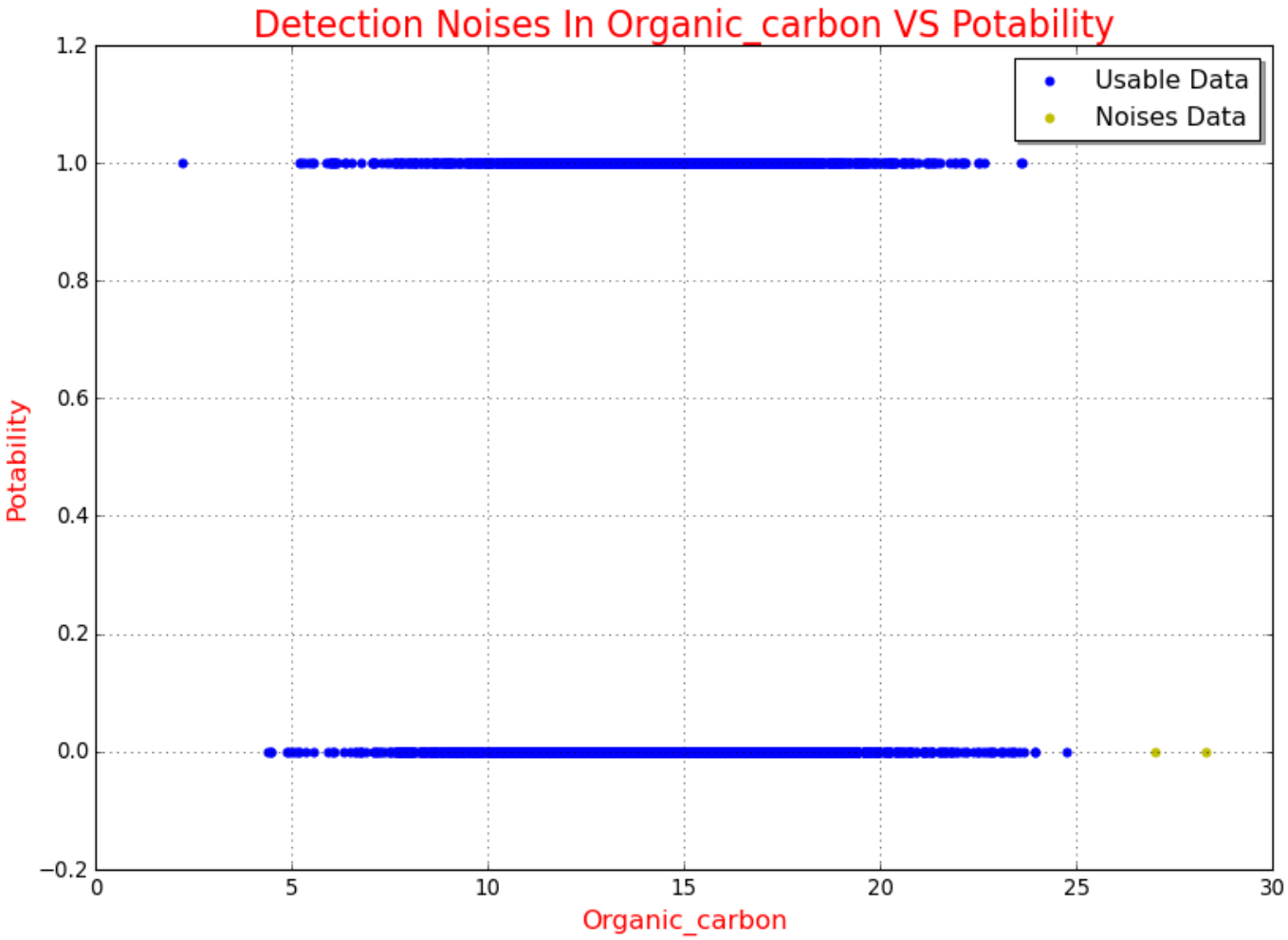

در شکل بالا دیتا های نویز با رنگ زرد مشخص شده اند. بر این اساس که کل مواد جامد در آب اگر بیشتر از ۵۰ هزار باشد قابل استفاده نیست.



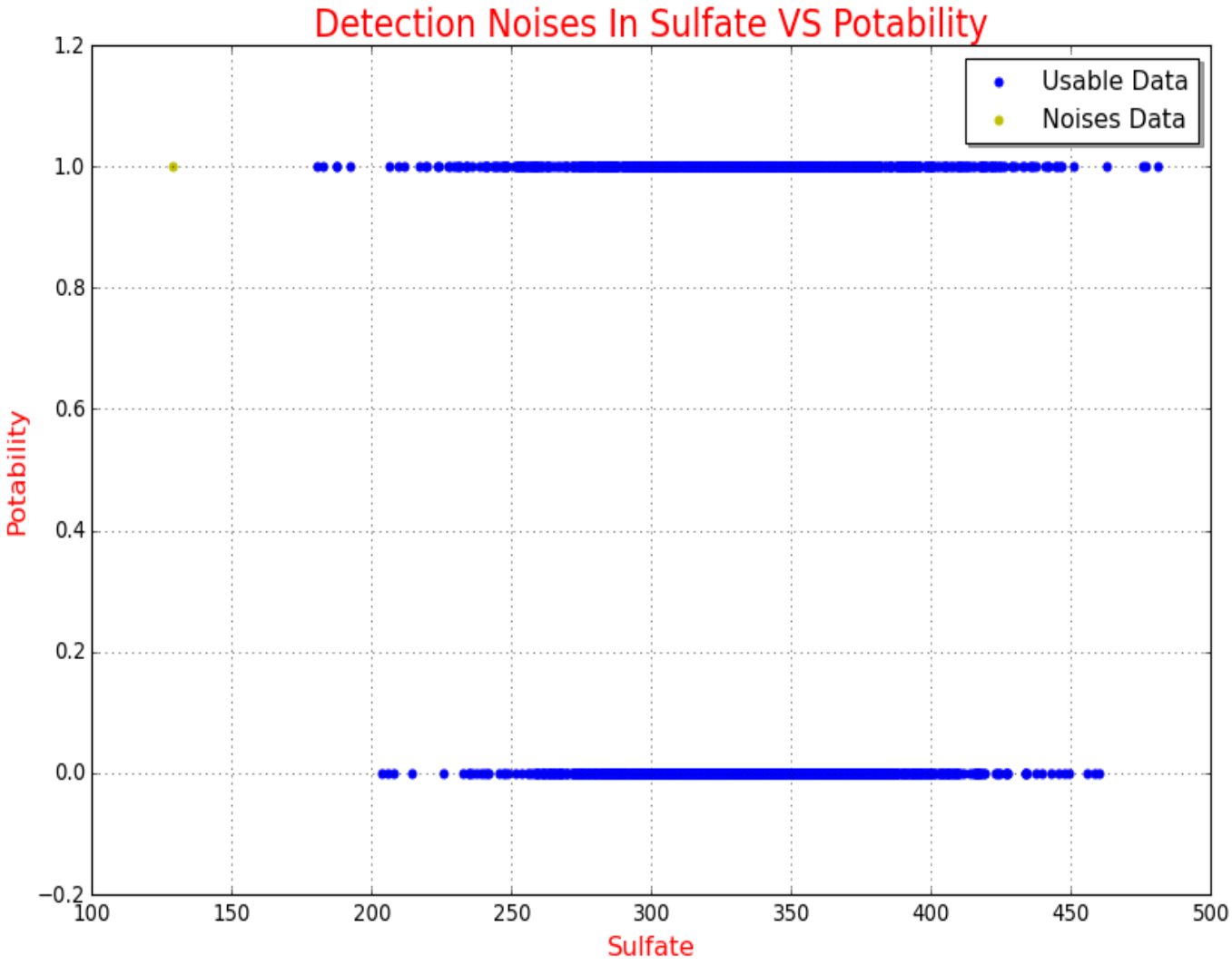

در شکل بالا دیتا های نویز با رنگ زرد مشخص شده اند. با توجه به دورافتادگی دیتاهای کمتر از ۲ در ستون کلرامین نسبت به سایر دیتاها تصمیم به حذف آنها گرفته شده است.



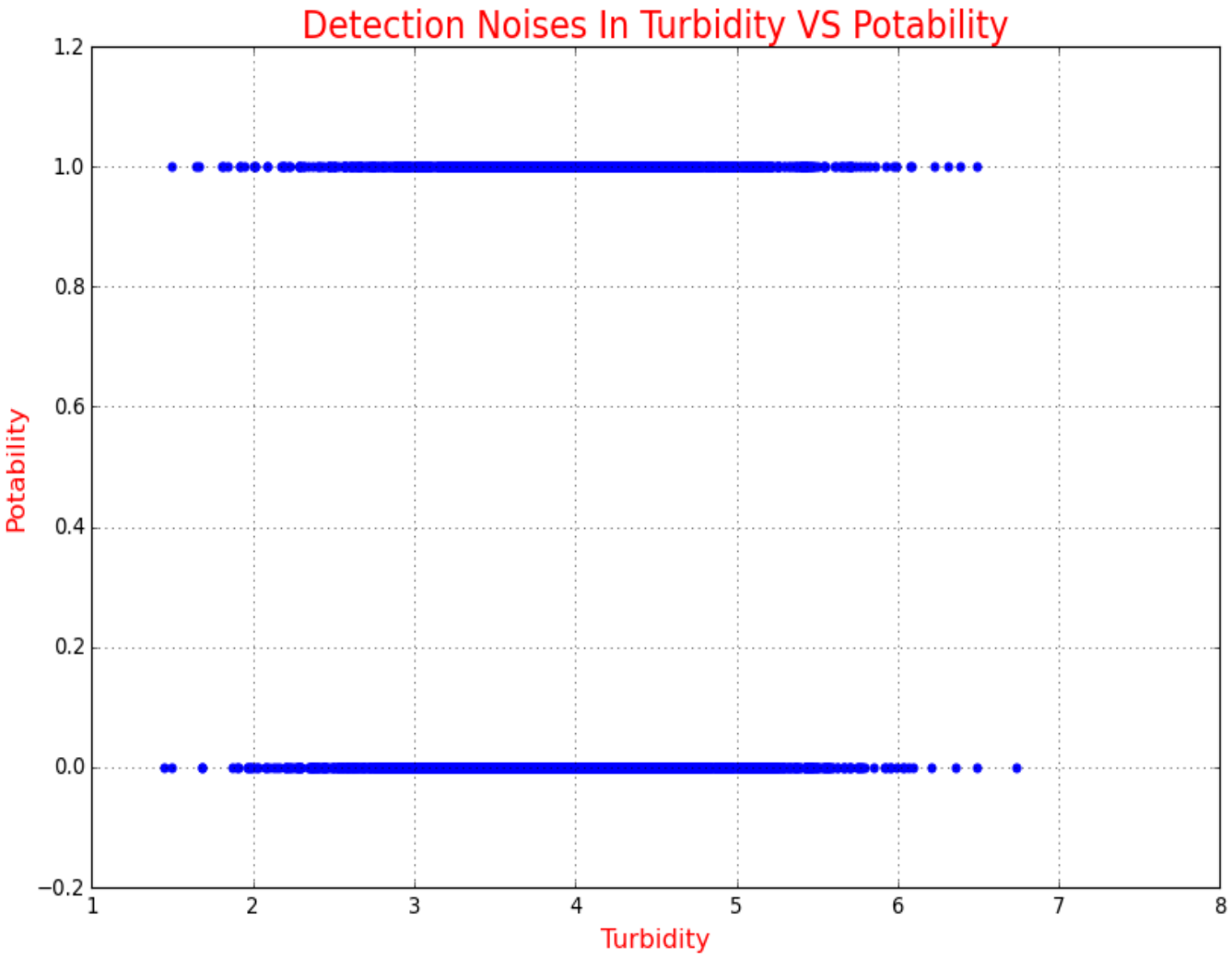

در شکل بالا دیتا های نویز با رنگ زرد مشخص شده اند. هدایت الکتریکی برای آب استخر نباید از ۲۰۰ کمتر باشد همچنین بیشتر از ۷۰۰ بودن میزان هدایت الکتریکی می تواند خطرناک باشد پس دیتاهای در این بازه به عنوان نویز شناخته می شوند.



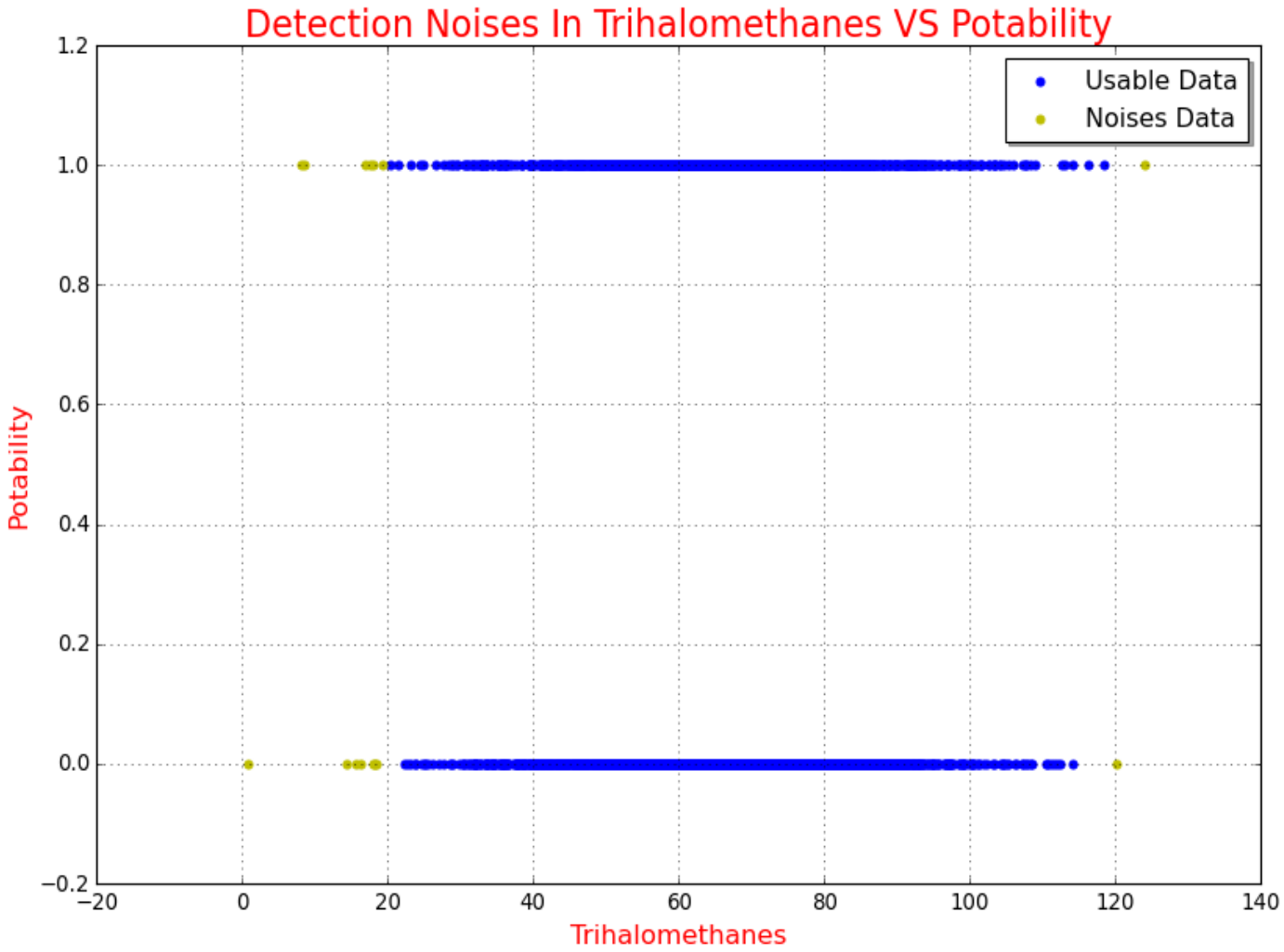

در شکل بالا دیتا های نویز با رنگ زرد مشخص شده اند. با توجه به فاصله زیاد دیتاهای بیشتر از ۲۵ppm نسبت به سایر دیتاها و مناسب نبودن دیتاهای بیشتر از ۲۰ برای آب استخر این دیتاها به عنوان نویز درنظر گرفته شده اند.




در شکل بالا دیتا های نویز با رنگ زرد مشخص شده اند. با توجه با فاصله زیاد دیتای نویز نسبت به سایر دیتاها این دیتا به عنوان دیتای نویز تشخیص داده شده است.




کل این محدوده برای میزان خاصیت انعکاس نور از سطح آب قابل قبول است . همچنین داده مشکوکی برای بررسی نویز بودن یا نبودن از لحاظ فاصله زیاد نسبت به سایر دیتاها وجود ندارد.




در شکل بالا دیتا های نویز با رنگ زرد مشخص شده اند. مقدار قابل قبول و مناسب تری هالومتان ها بر حسب میکروگرم در لیتر مقداری بین ۲۰ تا ۱۲۰ می باشد که هرچه این بازه کوچکتر باشد آب مصرفی برای استخر مناسب تر است.

## تشخیص و حذف داده های گمشده :

توضیحات آماری مختصری از فیچرهای دیتاست در شکل زیر آمده است. با توجه به سطر count می توان متوجه شد که در بعضی از ستون ها داده گمشده وجود دارد. جهت اطمینان از دستور Isnull().sum() نیز برای تشخیص داده های گمشده در ستون ها استفاده می شود.

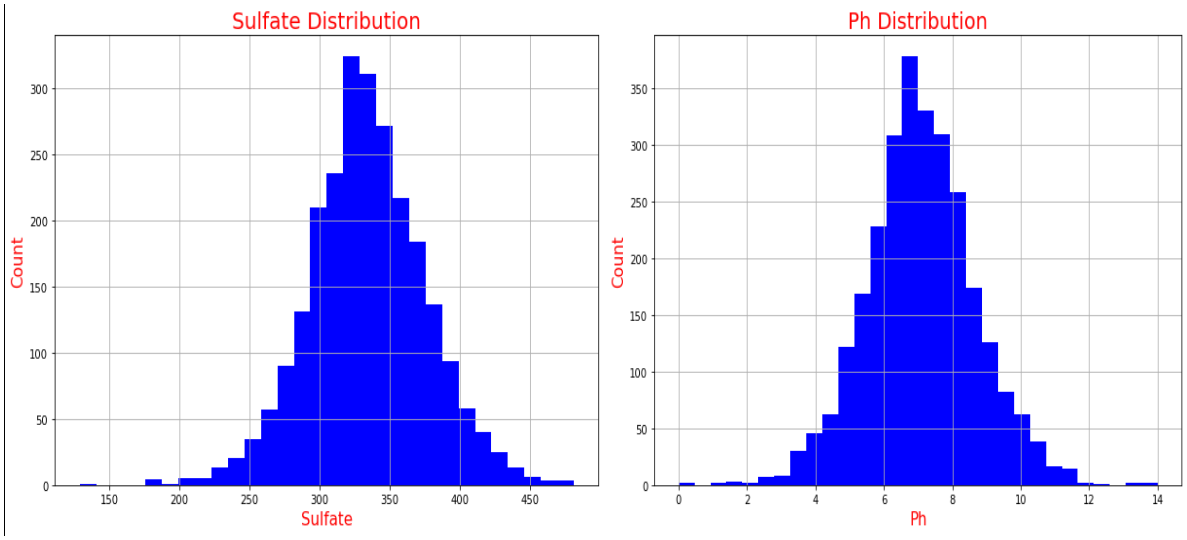
| | ph | Hardness | Solids | Chloramines | Sulfate | Conductivity | Organic_carbon | Trihalomethanes | Turbidity | Potability |
|---|---|---|---|---|---|---|---|---|---|---|
| count | 2785.000000 | 3276.000000 | 3276.000000 | 3276.000000 | 2495.000000 | 3276.000000 | 3276.000000 | 3114.000000 | 3276.000000 | 3276.000000 |
| mean | 7.080795 | 196.369496 | 22014.092526 | 7.122277 | 333.775777 | 426.205111 | 14.284970 | 66.396293 | 3.966786 | 0.390110 |
| std | 1.594320 | 32.879761 | 8768.570828 | 1.583085 | 41.416840 | 80.824064 | 3.308162 | 16.175008 | 0.780382 | 0.487849 |
| min | 0.000000 | 47.432000 | 320.942611 | 0.352000 | 129.000000 | 181.483754 | 2.200000 | 0.738000 | 1.450000 | 0.000000 |
| 25% | 6.093092 | 176.850538 | 15666.690297 | 6.127421 | 307.699498 | 365.734414 | 12.065801 | 55.844536 | 3.439711 | 0.000000 |
| 50% | 7.036752 | 196.967627 | 20927.833607 | 7.130299 | 333.073546 | 421.884968 | 14.218338 | 66.622485 | 3.955028 | 0.000000 |
| 75% | 8.062066 | 216.667456 | 27332.762127 | 8.114887 | 359.950170 | 481.792304 | 16.557652 | 77.337473 | 4.500320 | 1.000000 |
| max | 14.000000 | 323.124000 | 61227.196008 | 13.127000 | 481.030642 | 753.342620 | 28.300000 | 124.000000 | 6.739000 | 1.000000 |

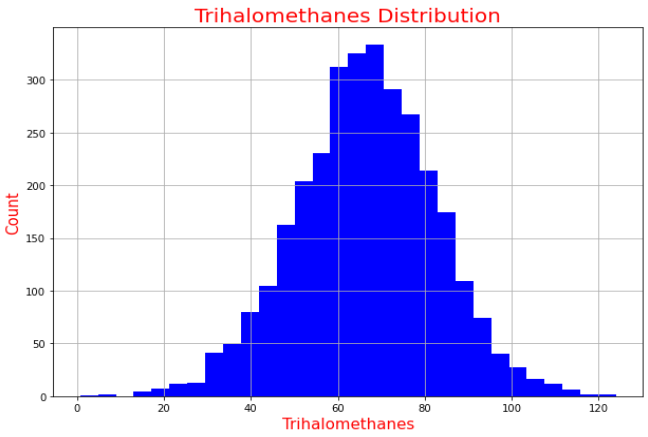
```
Show the missing values of DataFrame:

 ph                491
Hardness             0
Solids               0
Chloramines          0
Sulfate            781
Conductivity         0
Organic_carbon       0
Trihalomethanes    162
Turbidity            0
Potability           0
dtype: int64
```

همانطور که مشخص است در ۳ ستون ph و Sulfate و Trihalomethanes داده گمشده وجود دارد. با توجه به تعداد زیاد داده های گمشده در هر ستون و اهمیت بالای ستون هایی مثل ph حذف کردن گزینه اول

نیست بلکه باید رفتار دیتا را برای جایگزینی مقادیر NAN بسنجیم و سپس براساس آن اقدام به پرکردن داده های گمشده کنیم.




شکل های بالا نمودار هیستوگرام ستون هایی هستند که دارای داده های گمشده می باشند با توجه به رفتار دیتا در نمودارهای بالا یکی از راه های مناسب برای پرکردن دیتاها از طریق روش اینترپوله کردن است.

## حذف نویزها :

حال نوبت به حذف نویزها است. بعد از تشخیص نویزها حذف کردن آن ها کار چندان مشکلی نیست و با چند خط کد اجرا میشود در زیر کد حذف و خروجی دیتاست بعد از اعمال تغییرات ( حذف داده های گمشده و نویزها ) نشان داده شده است.

In [181]:
```
df2 = df[df['Hardness'] > 100]
df3 = df2[df2['Solids'] < 50000]
df4 = df3[(df3['Conductivity'] < 700) & (df3['Conductivity'] > 200)]
df5 = df4[df4['Organic_carbon'] < 25]
df6 = df5[df5['Sulfate'] > 150]
df7 = df6[(df6['Trihalomethanes'] > 20) & ( df6['Trihalomethanes'] < 120)]
df7 = df7.reset_index(drop=True)
df7
```

Out[181]:

| | ph | Hardness | Solids | Chloramines | Sulfate | Conductivity | Organic_carbon | Trihalomethanes | Turbidity | Potability |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 3.716080 | 204.890455 | 20791.318981 | 7.300212 | 368.516441 | 564.308654 | 10.379783 | 86.990970 | 2.963135 | 0 |
| **1** | 3.716080 | 129.422921 | 18630.057858 | 6.635246 | 364.639673 | 592.885359 | 15.180013 | 56.329076 | 4.500656 | 0 |
| **2** | 8.099124 | 224.236259 | 19909.541732 | 9.275884 | 360.762904 | 418.606213 | 16.868637 | 66.420093 | 3.055934 | 0 |
| **3** | 8.316766 | 214.373394 | 22018.417441 | 8.059332 | 356.886136 | 363.266516 | 18.436524 | 100.341674 | 4.628771 | 0 |
| **4** | 9.092223 | 181.101509 | 17978.986339 | 6.546600 | 310.135738 | 398.410813 | 11.558279 | 31.997993 | 4.075075 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **3227** | 4.668102 | 193.681735 | 47580.991603 | 7.166639 | 359.948574 | 526.424171 | 13.894419 | 66.687695 | 4.435821 | 1 |
| **3228** | 7.808856 | 193.553212 | 17329.802160 | 8.061362 | 359.948574 | 392.449580 | 19.903225 | 68.266548 | 2.798243 | 1 |
| **3229** | 9.419510 | 175.762646 | 33155.578218 | 7.350233 | 359.948574 | 432.044783 | 11.039070 | 69.845400 | 3.298875 | 1 |
| **3230** | 5.126763 | 230.603758 | 11983.869376 | 6.303357 | 359.948574 | 402.883113 | 11.168946 | 77.488213 | 4.708658 | 1 |
| **3231** | 7.874671 | 195.102299 | 17404.177061 | 7.509306 | 359.948574 | 327.459760 | 16.140368 | 78.698446 | 2.309149 | 1 |

3232 rows × 10 columns

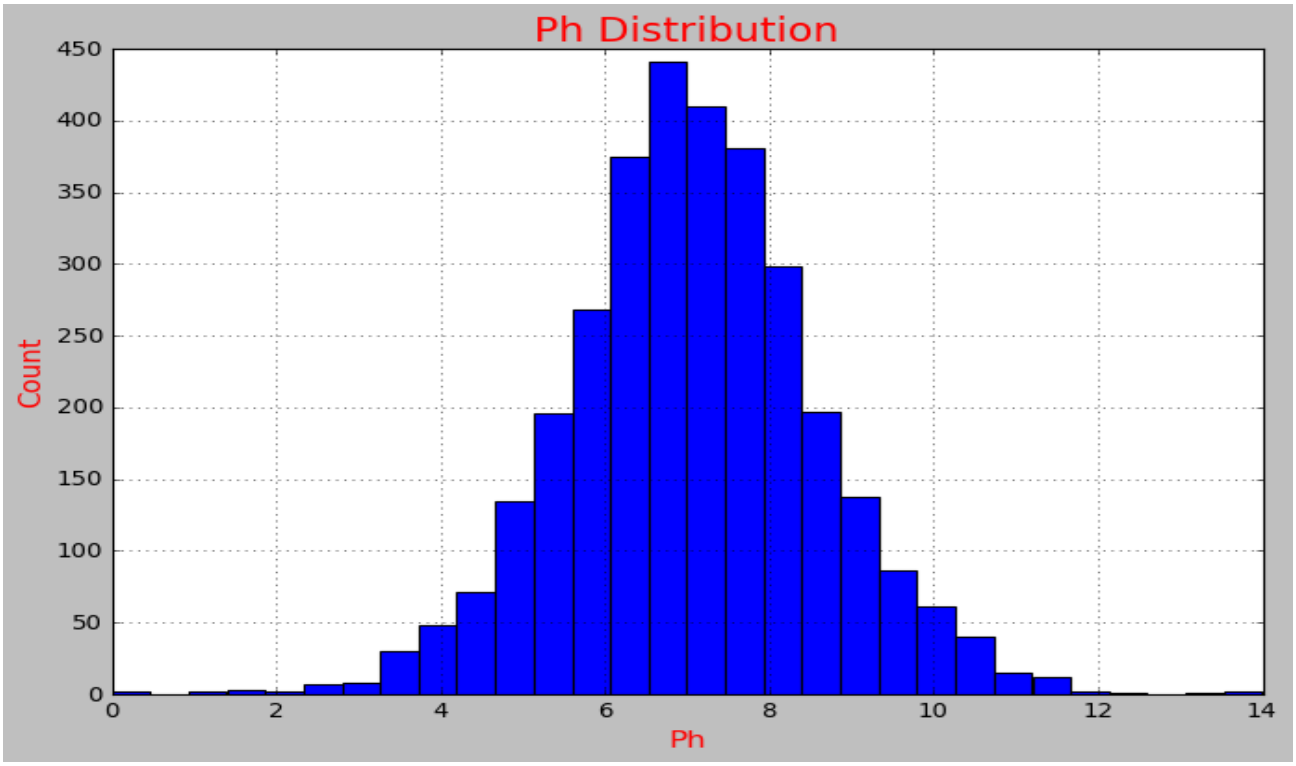
# مرحله سوم : بازپردازش

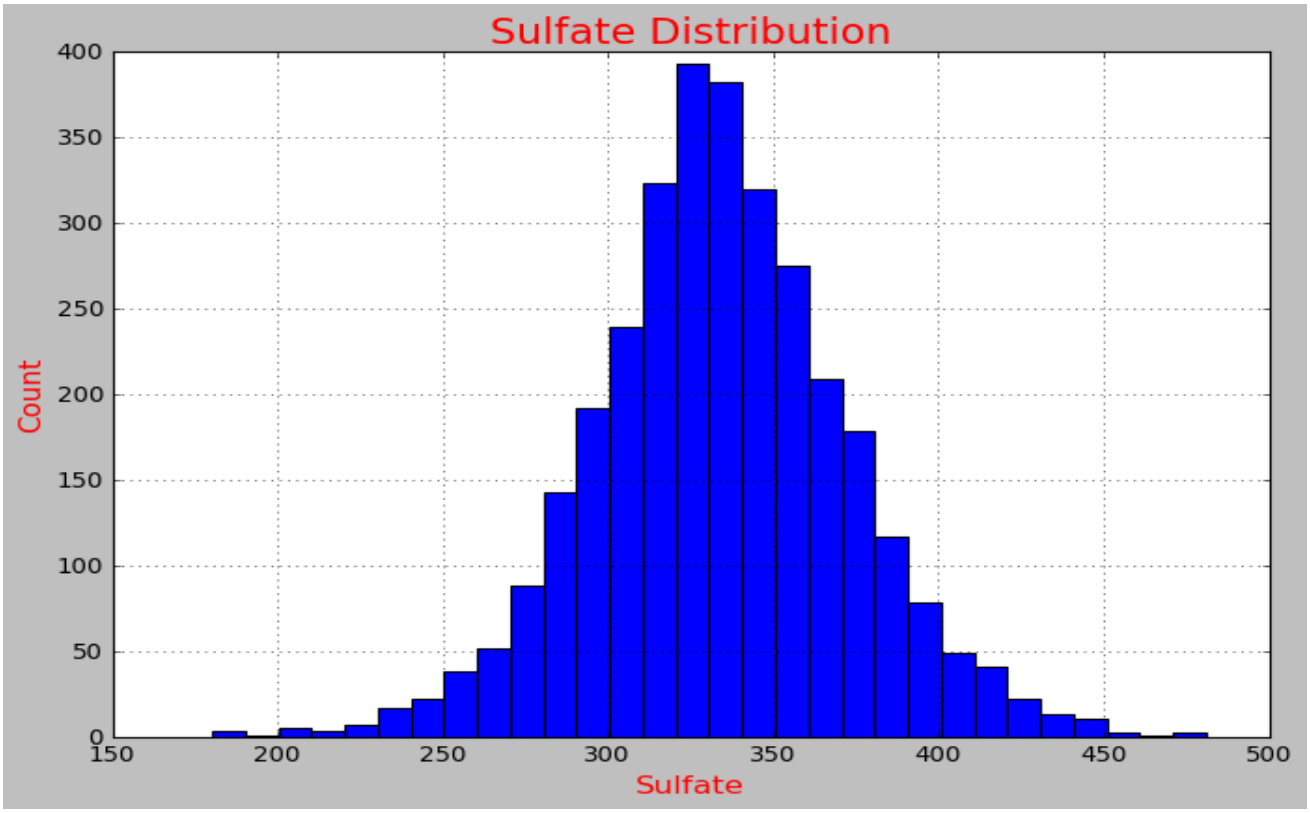
پس از پیش پردازش و پاکسازی دیتا از هرگونه خطا، علم خوبی از دیتا بدست آورده ایم. اما این علم بسیار پراکنده و نامنظم است. برای همین در این مرحله ما اطلاعات خود را جمع بندی خواهیم کرد.

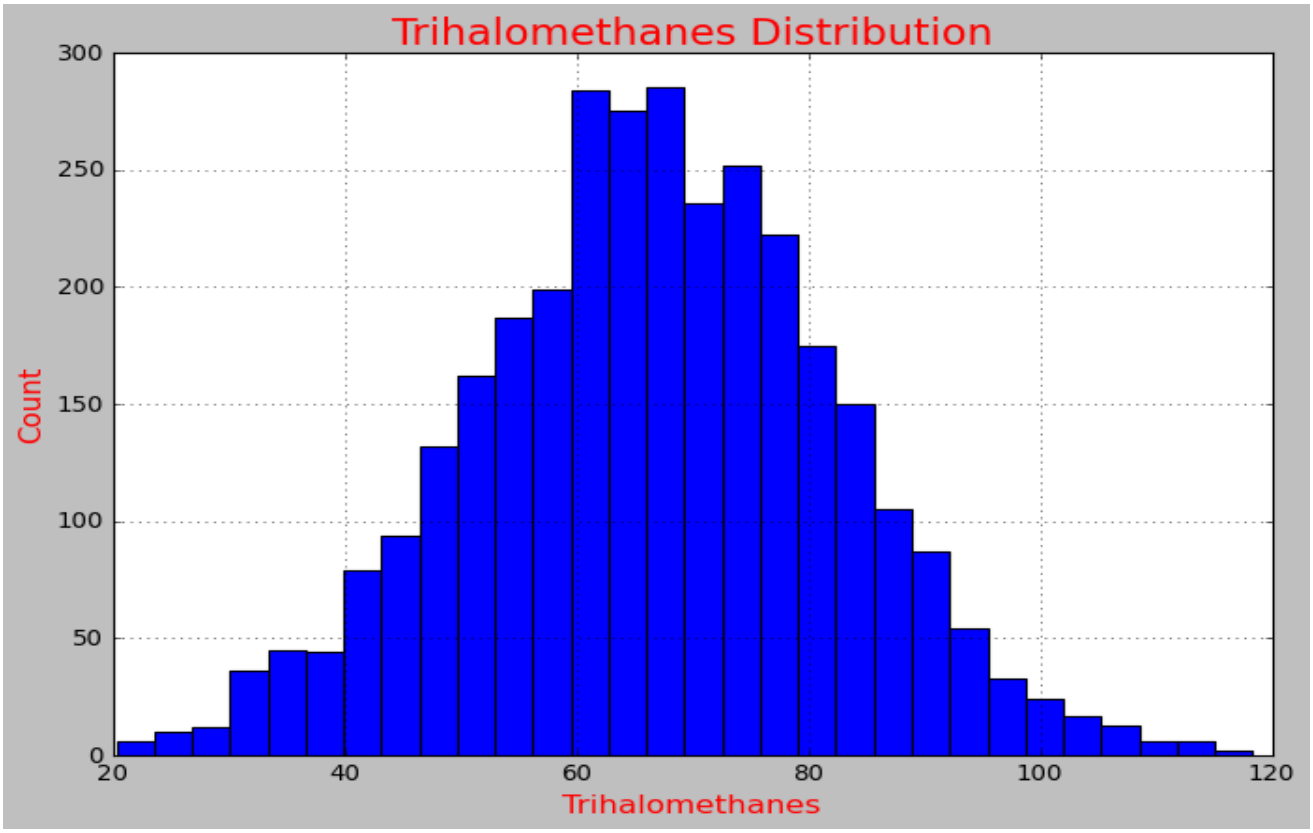
بعد از بررسی ها و تغییراتی که در مراحل قبلی بر روی دیتاست اعمال شد این نتیجه حاصل شد که تعدادی از دیتاها مناسب برای ادامه کار با دیتاست نیستند و نویز تشخیص داده شدند بنابراین آنها را حذف کرده ایم همچنین بعضی از ستون ها شامل تعداد زیادی دیتاهای گمشده بودند که با توجه به رفتار فیچرها داده های گمشده را با دیتاهای مناسب پر کردیم با توجه به اهمیت فیچرها در دیتاست نمی توانستیم آنها را از دیتاست حذف کنیم. پس از روشی آنها را پر کردیم که کمترین انحراف در توزیع آن دیده شود و از روش اینترپوله کردن برای این منظور استفاده کردیم.

| | ph | Hardness | Solids | Chloramines | Sulfate | Conductivity | Organic_carbon | Trihalomethanes | Turbidity | Potability |
|---|---|---|---|---|---|---|---|---|---|---|
| count | 3232.000000 | 3232.000000 | 3232.000000 | 3232.000000 | 3232.000000 | 3232.000000 | 3232.000000 | 3232.000000 | 3232.000000 | 3232.000000 |
| mean | 7.059333 | 196.680436 | 21880.208672 | 7.122904 | 333.869045 | 426.088116 | 14.283334 | 66.613548 | 3.967114 | 0.387376 |
| std | 1.525370 | 32.265223 | 8550.196328 | 1.578316 | 38.607001 | 80.556445 | 3.296610 | 15.615605 | 0.780866 | 0.487226 |
| min | 0.000000 | 100.457615 | 320.942611 | 0.352000 | 180.206746 | 201.619737 | 2.200000 | 20.337753 | 1.450000 | 0.000000 |
| 25% | 6.113385 | 177.091694 | 15597.348996 | 6.128831 | 310.378554 | 365.672262 | 12.064863 | 56.299562 | 3.439540 | 0.000000 |
| 50% | 7.031762 | 197.063450 | 20882.160702 | 7.128299 | 333.073455 | 421.926811 | 14.220645 | 66.715011 | 3.954025 | 0.000000 |
| 75% | 7.985559 | 216.667456 | 27250.471317 | 8.109544 | 357.447158 | 481.376537 | 16.560201 | 77.228284 | 4.500320 | 1.000000 |
| max | 14.000000 | 323.124000 | 49456.587108 | 13.127000 | 481.030642 | 695.369528 | 24.755392 | 118.357275 | 6.739000 | 1.000000 |

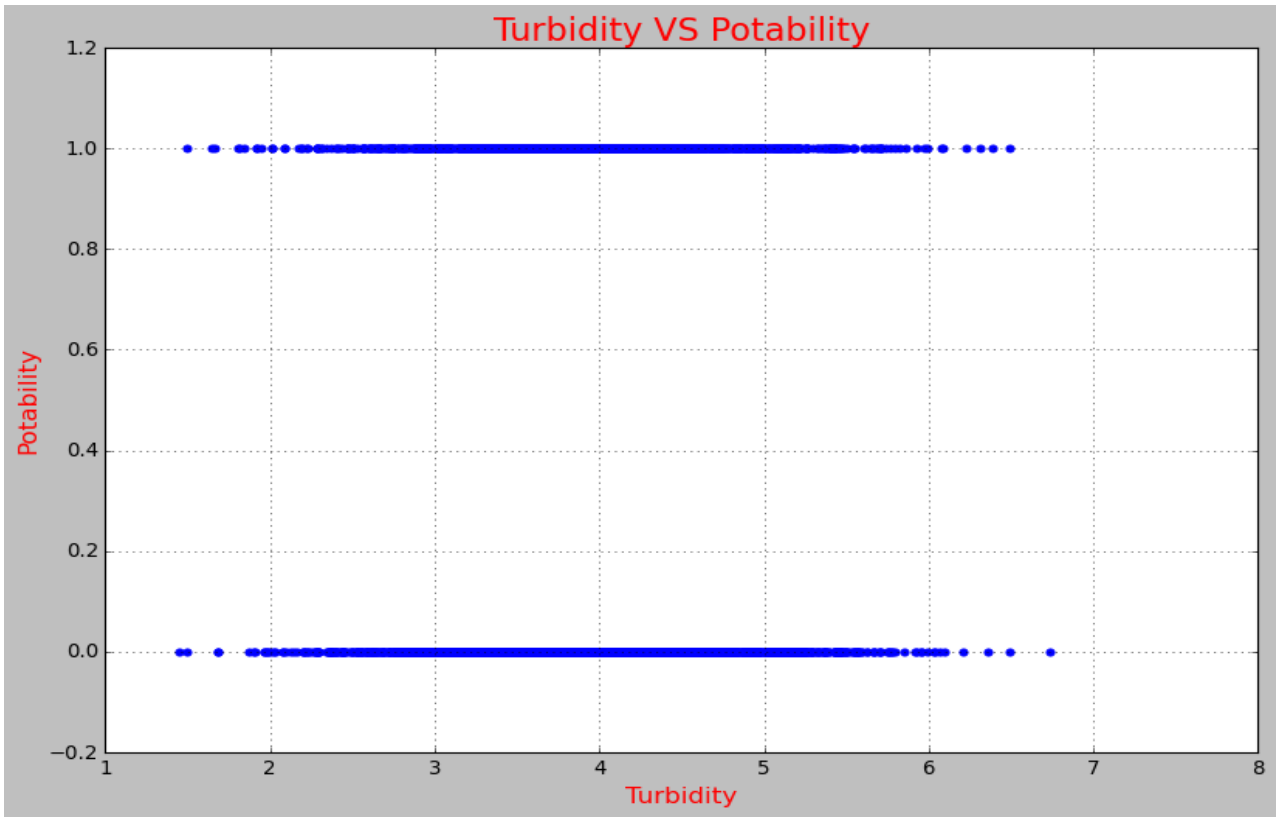
شکل بالا اطلاعات آماری دیتاست نهایی و پاکسازی شده و آماده برای مدسازی ماشین لرنینگ را نشان می دهد.

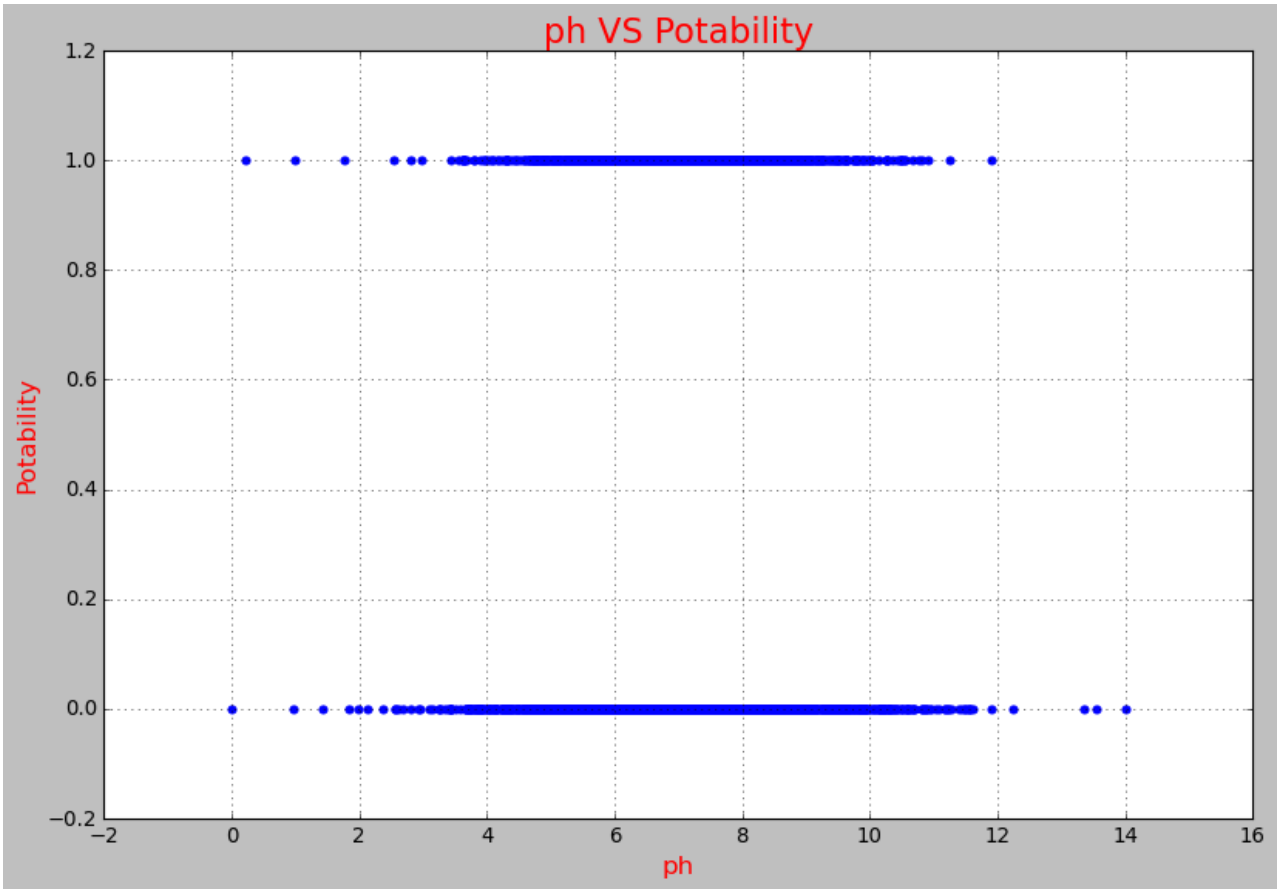

Ph Distribution
Count
Ph
0
50
100
150
200
250
300
350
400
450
0
2
4
6
8
10
12
14

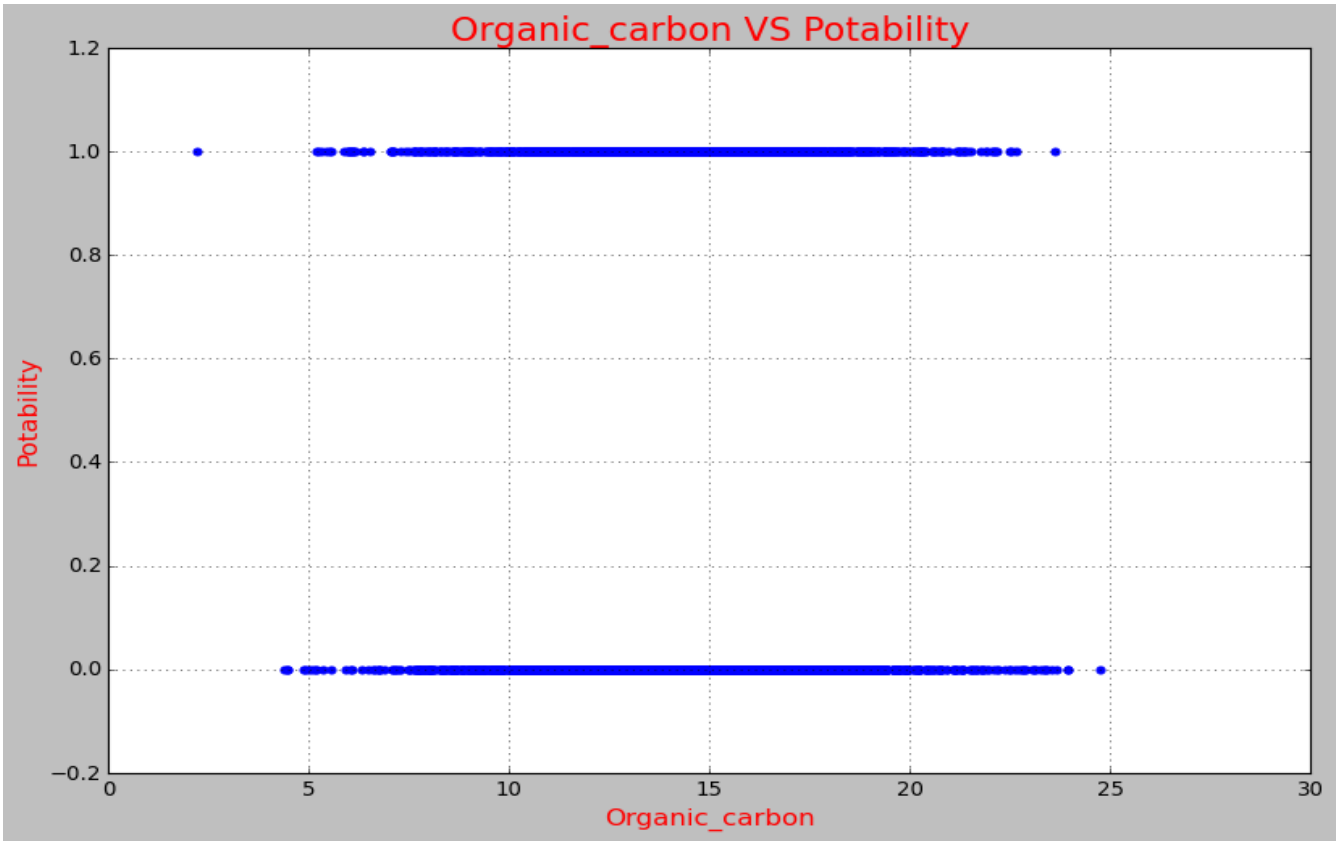

Sulfate Distribution
Count
Sulfate
0
50
100
150
200
250
300
350
400
150
200
250
300
350
400
450
500

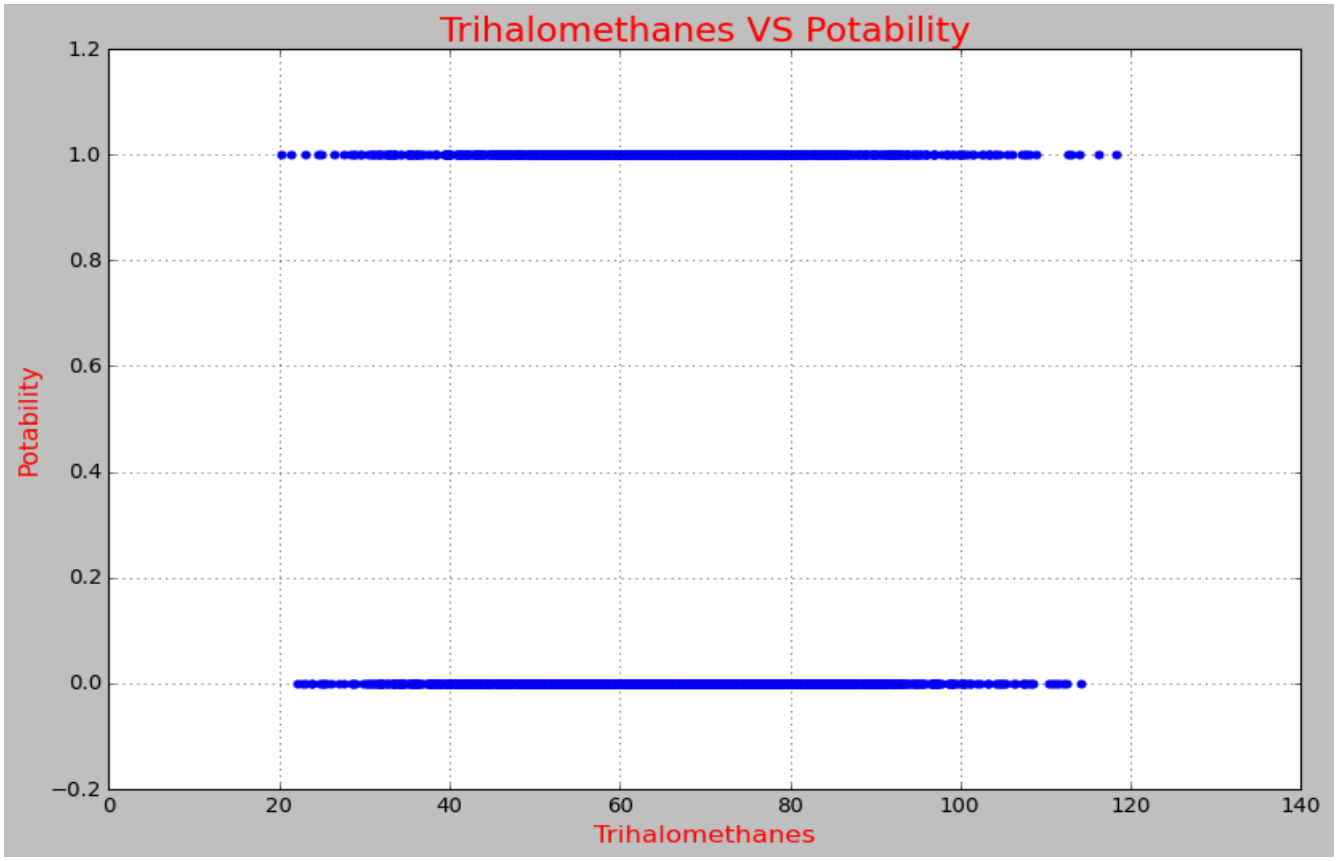


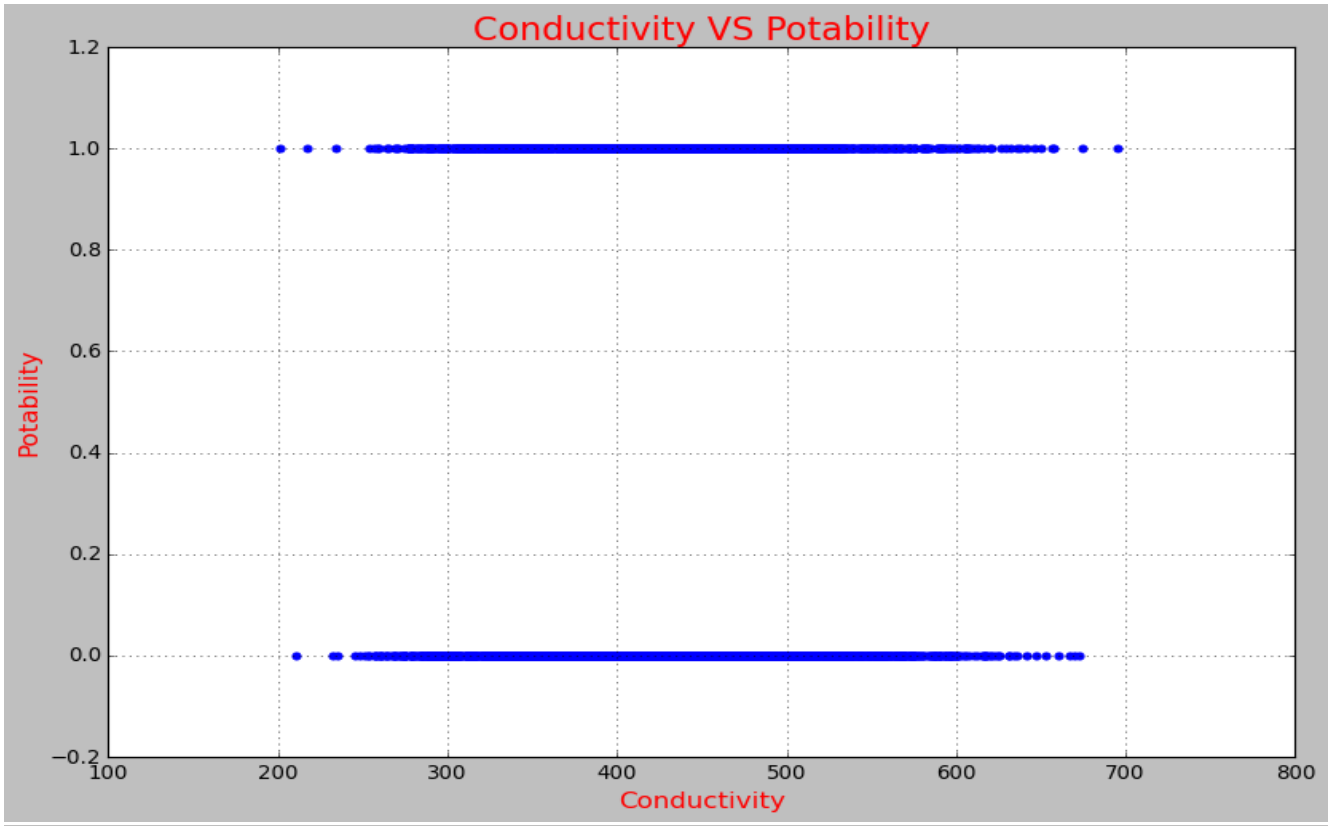

رفتار فیچرهایی از دیتاست که شامل داده های گمشده بودند بعد از اعمال تغییرات و پر کردن داده های گمشده با مقادیر مناسب از طریق روش اینترپوله کردن به شکل بالا می باشد.

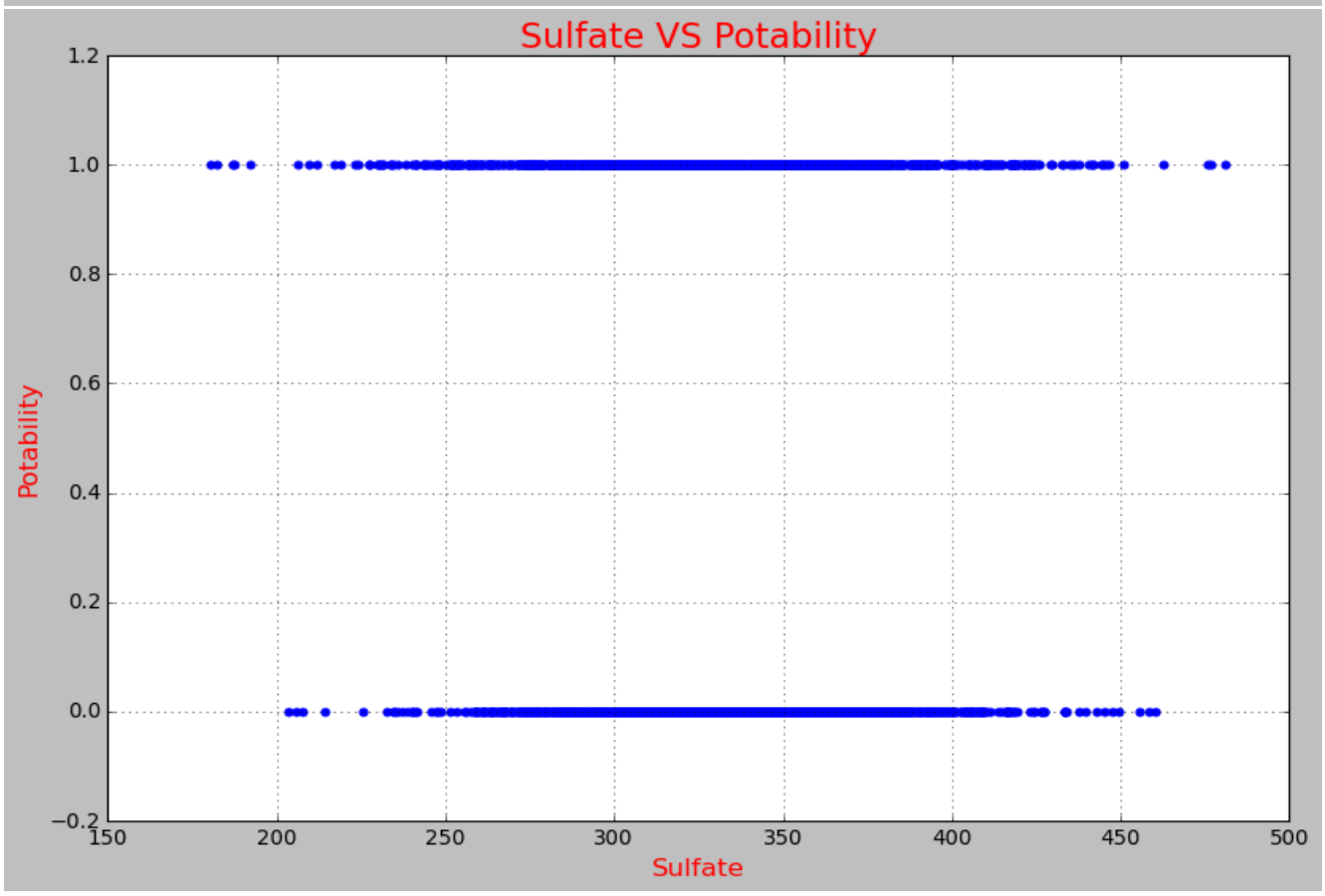

Turbidity VS Potability
Potability
Turbidity
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
1
2
3
4
5
6
7
8

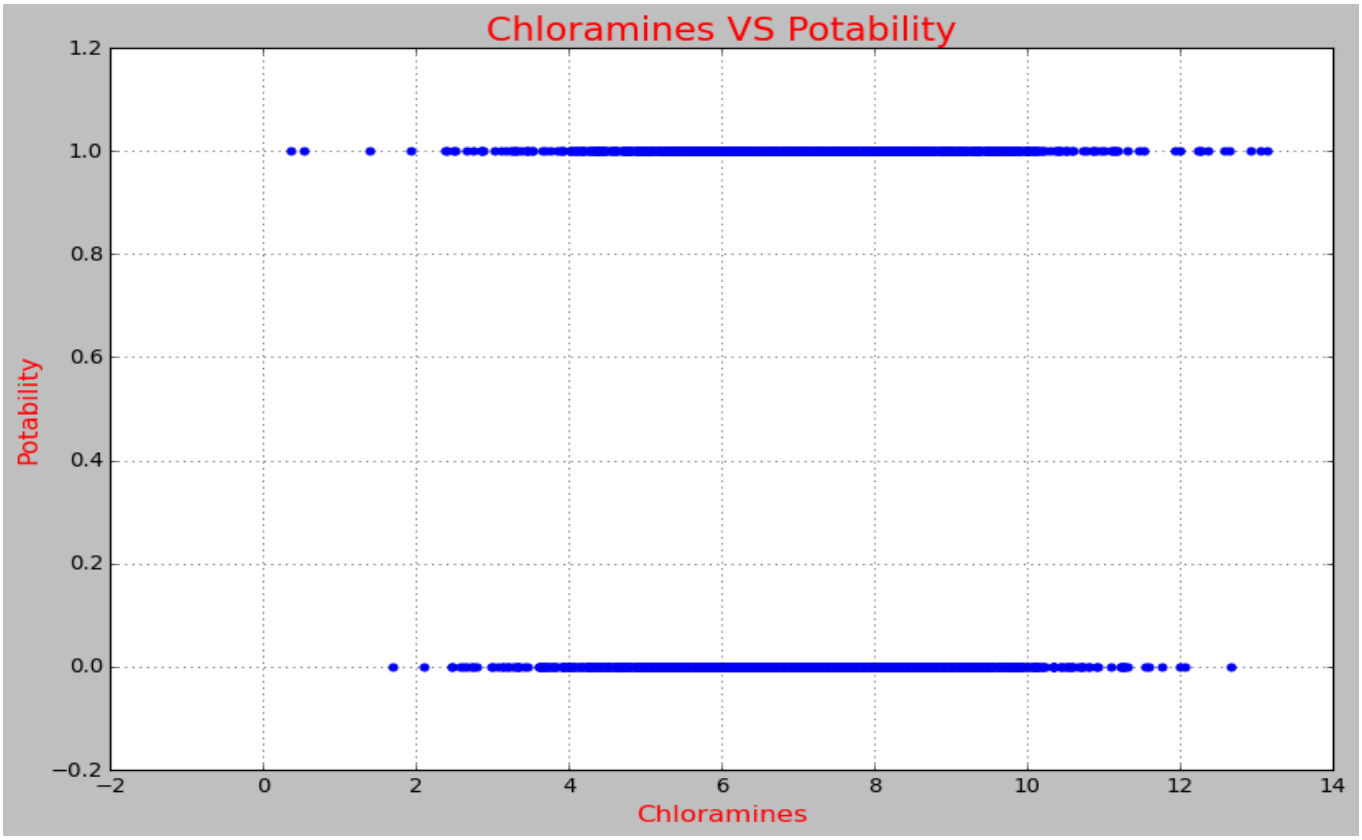

ph VS Potability
Potability
ph
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−2
0
2
4
6
8
10
12
14
16

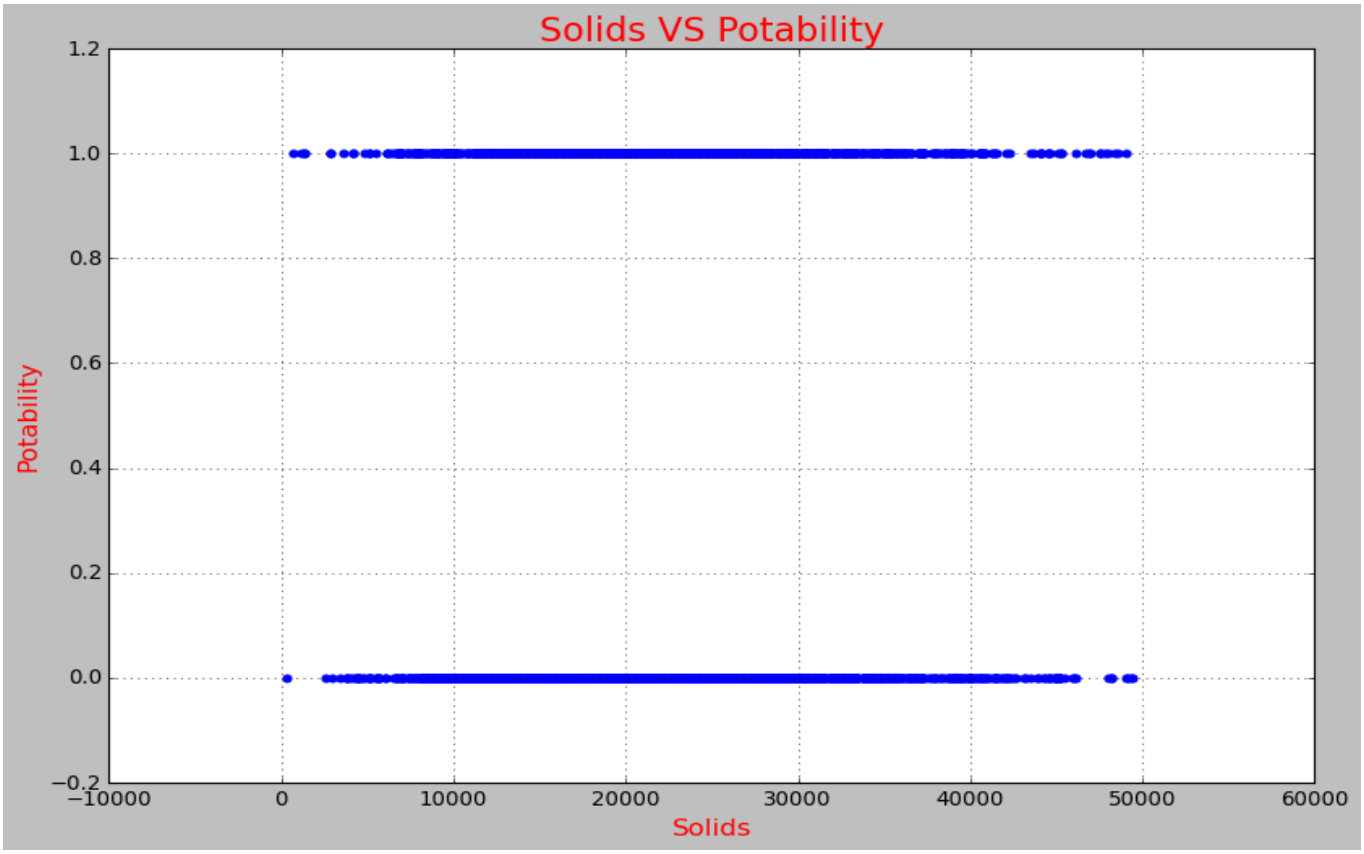

Organic_carbon VS Potability
Potability
Organic_carbon
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
5
10
15
20
25
30

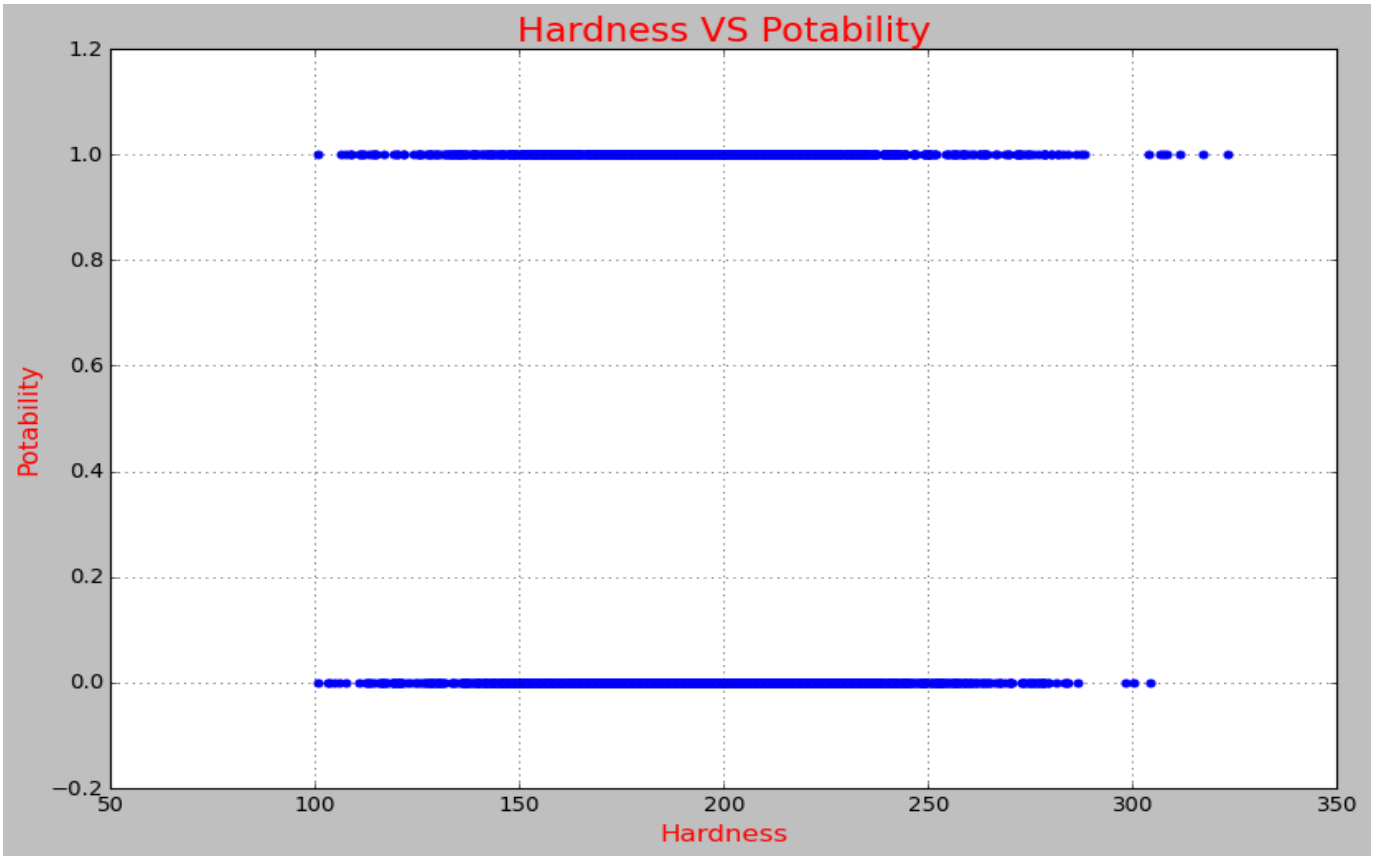

Trihalomethanes VS Potability
Potability
Trihalomethanes
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
20
40
60
80
100
120
140


Conductivity VS Potability
Potability
Conductivity
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
100
200
300
400
500
600
700
800


Sulfate VS Potability
Potability
Sulfate
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
150
200
250
300
350
400
450
500


Chloramines VS Potability
Potability
Chloramines
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−2
0
2
4
6
8
10
12
14


Solids VS Potability
Potability
Solids
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−10000
0
10000
20000
30000
40000
50000
60000




## حدس پترن :

با توجه به اینکه y یا تارگت در مسئله ما وجود دارد و به ازای هر ورودی یک خروجی مشخص داریم که در طبقات ۱ یا ۰ قرار میگیرد می توان نتیجه گرفت که مسئله ما از نوع یادگیری باناظر است و طبیعتا باید از الگوریتم های این نوع یادگیری برای مدلسازی ماشین لرنینگ دیتاست خود استفاده کنیم.