Arsenic data

Austin Martinez

2024-06-13

## Background

### Overview Statement

This dataset that is provided by the U.S. Geological Survey (USGS) was compiled and published by Melissa A. Lombard in 2021. This data spans from 1970 to 2013. This data was originally sourced to aid in the development of models that estimate the probability of private well arsenic concentrations exceeding various thresholds (1, 5, and 10 micrograms per liter) throughout the conterminous United States.

The purpose of retrieving this dataset is to develop a model that uses this data to predict which areas of the Colorado Plateau are most at risk for underground water contamination by arsenic.

## Geologic Predictor Variables

#provide summaries of geologic predictor variables here, examples may include rocks, soils and other geologic factors

* **rt\_carb**: Generalized lithology classified as CARB
* **rt\_clast\_c**: Generalized lithology classified as CLAST\_C
* **rt\_clast\_f**: Generalized lithology classified as CLAST\_F
* **rt\_clast\_u**: Generalized lithology classified as CLAST\_U
* **rt\_meta**: Generalized lithology classified as META
* **rt\_plut\_qtz**: Generalized lithology classified as PLUT\_QTZ
* **D3**: Devonian; stratified sequence
* **De**: Devonian eugeosynclinal deposits
* **DSe**: New England; Devonian and Silurian eugeosynclinal deposits
* **Kg**: Plutonic and intrusive rocks; Cretaceous granitic rocks
* **O2**: Ordovician; stratified sequence
* **Oe**: Ordovician, eugeosynclinal
* **PP4**: Upper Pennsylvanian; stratified sequence
* **Pzg1**: Lower Paleozoic granitic rocks
* **Pzg2**: Middle Paleozoic granitic rocks
* **Q**: Quaternary; stratified sequence
* **Qv**: Quaternary volcanic rocks
* **S2**: Silurian; stratified sequence
* **Se**: Silurian eugeosynclinal deposits
* **Te2**: Tertiary; Eocene; stratified sequence
* **Tm**: Tertiary; stratified sequence
* **Tmc**: Tertiary; Miocene continental deposits
* **Toc**: Tertiary; Oligocene continental deposits
* **Tp**: Tertiary; stratified sequence
* **Tpc**: Tertiary; Pliocene continental deposits
* **Tpv**: Tertiary; Pliocene volcanic rocks
* **Tr**: Triassic; stratified sequence
* **Txc**: Tertiary; Paleocene continental deposits
* **uK2**: Upper Cretaceous Austin and Eagle Ford Groups
* **uK3**: Upper Cretaceous Taylor Group
* **uK3b**: Upper Cretaceous Taylor Group
* **uK4**: Upper Cretaceous Navarro Group
* **uPz**: Upper Paleozoic; stratified sequence
* **Wgn**: PreCambrian; metamorphic rocks
* **as\_c**: Arsenic concentration in the soil C horizon
* **be\_c**: Beryllium concentration in the soil C horizon
* **bi\_c**: Bismuth concentration in the soil C horizon
* **inorgc\_c**: Inorganic carbon concentration in the soil C horizon
* **mo\_c**: Molybdenum concentration in the soil C horizon
* **ni\_c**: Nickel concentration in the soil C horizon
* **orgc\_c**: Phosphorus concentration in the soil C horizon
* **p\_c**: Organic carbon concentration in the soil C horizon
* **sb\_c**: Antimony concentration in the soil C horizon
* **se\_c**: Selenium concentration in the soil C horizon

## Hydrology Predictor Variables

#provide summaries of hydrologic predictor variables here, examples may include depth to groundwater, lateral hydrologic position, stream flow, etc, acquifer characteristics

* **LP2**: Lateral position for fourth order streams
* **PRMS8110Re**: Average annual recharge to groundwater
* **LP2**: Lateral position for fourth order streams
* **LP4**: Lateral position for second order streams
* **LP6**: Lateral position for sixth order streams
* **BFI**: Mean annual base flow index
* **uc\_999**: Water
* **Percent\_Ti**: County estimates of U.S. tile drainage

## Climate Predictor Variables

#provide summaries of hydrologic predictor variables here, examples may include precipitation, temperature, evapotranspiration

* **DMppt8110**: Average annual precipitation from 1981 to 2010

## Ecological Region Variables (idk if you wanna call this land cover Predictor Variables)

* **na\_10.1**: Cold deserts ecoregion
* **na\_11.1**: Mediterranean California ecoregion
* **na\_12.1**: Western Sierra Madre piedmont ecoregion
* **na\_13.1**: Upper Gila Mountains ecoregion
* **na\_6.2**: Western Cordillera ecoregion
* **na\_7.1**: Marine West Coast forest ecoregion
* **na\_8.1**: Mixed wood plains ecoregion
* **na\_8.2**: Central USA plains ecoregion
* **na\_9.2**: Temperate prairies ecoregion
* **na\_9.3**: West-Central semi-arid prairies ecoregion
* **na\_9.4**: South Central semi-arid prairies ecoregion
* **na\_9.5**: Texas-Louisiana coastal plain ecoregion
* **na\_9.6**: Tamaulipas-Texas semiarid plain ecoregion

## Soil and Sediment Variables

* **uc\_12**: Alluvial sediments
* **uc\_221**: Coastal zone sediments
* **uc\_312**: Eolian sediments
* **uc\_321**: Eolian sediments
* **uc\_322**: Eolian sediments
* **uc\_331**: Eolian sediments
* **uc\_421**: Glacial till sediments
* **uc\_422**: Glacial till sediments
* **uc\_431**: Glacial till sediments
* **uc\_451**: Glaciofluvial ice-contact sediments
* **uc\_620**: Colluvial and alluvial sediments
* **uc\_811**: Proglacial sediments
* **uc\_812**: Proglacial sediments
* **uc\_821**: Proglacial sediments
* **uc\_822**: Proglacial sediments
* **uc\_910**: Residual materials developed in igneous and metamorphic rocks
* **uc\_920**: Residual materials developed in sedimentary rocks
* **uc\_940**: Residual materials developed in carbonate rocks
* **uc\_960**: Residual materials developed in bedrock
* **uc\_970**: Residual materials developed in bedrock

## Output Variables

* **As3Cat**: Code to assign observed arsenic concentration to categorical variable
* **bas1**: Code to assign observed arsenic concentration to binary variable
* **bas10**: Code to assign observed arsenic concentration to binary variable
* **bas5**: Code to assign observed arsenic concentration to binary variable
* **spl1**: Code to indicate if the location was used in the boosted regression tree As 1 model training or testing dataset
* **spl10**: Code to indicate if the location was used in the boosted regression tree As 10 model training or testing dataset
* **spl3cat**: Code to indicate if the location was used in the random forest model training or testing dataset
* **spl5**: Code to indicate if the location was used in the boosted regression tree As 5 model training or testing dataset

# U.S. Geological Survey (USGS) Dataset

#setwd("/Users/austinmartinez/Documents/GitHub/coPlateauWaterQuality/01\_data")  
setwd("C:/Users/jhoover/Documents/GitHub/coPlateauWaterQuality/01\_data")  
library(kableExtra)

## Warning: package 'kableExtra' was built under R version 4.3.3

arsenicdata = read.csv("Model\_input\_definitions 1.csv")  
kable(arsenicdata)

| Predictor\_name | Description | Category\_levels | Categorical.observation.amounts | Data\_type | Units | Missing\_values | Time\_period | Source | Mean | X25..Quartile | X75..Quartile | Standard.Deviation | Max | Min |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| rt\_carb | Generalized lithology classified as CARB: sediments and sedimentary rocks. Carbonate rocks such as limestone and dolostone. Generally, any rock including any minor carbonate lithology is included in this group. Some special cases exist where carbonates are also identified based on LITH62MINO field in Schweitzer, Peter N., 2011, Combined geologic map data for the conterminous US derived from the USGS state geologic map compilation. | 1 (present) ; 0 (absent) | 0 = 17306 1 = 3144 | binary | NA | “NA” | NA | Anning and Ator (2017) | 0.1537 | 0.0000000 | 0.0000000 | 0.3607090 | 1 | 0 |
| rt\_clast\_c | Generalized lithology classified as CLAST\_C: sediments and sedimentary rocks. Clastic sediments/rocks primarily made of sands, gravels, cobbles, or larger clasts. | 1 (present) ; 0 (absent) | 0 = 15343 1 = 5107 | binary | NA | “NA” | NA | Anning and Ator (2017) | 0.2497 | 0.0000000 | 0.0000000 | 0.4328679 | 1 | 0 |
| rt\_clast\_f | Generalized lithology classified as CLAST\_F: sediments and sedimentary rocks. Clastic sediments/rocks primarily made of fine-grained materials such as shale, siltstone, claystone, mudstone, etc. | 1 (present) ; 0 (absent) | 0 = 16422 1 = 4028 | binary | NA | “NA” | NA | Anning and Ator (2017) | 0.197 | 0.0000000 | 0.0000000 | 0.3977178 | 1 | 0 |
| rt\_clast\_u | Generalized lithology classified as CLAST\_U: sediments and sedimentary rocks. Clastic sediments/rocks of unknown or highly variable clast sizes. | 1 (present) ; 0 (absent) | 0 = 16287 1 = 4163 | binary | NA | “NA” | NA | Anning and Ator (2017) | 0.2036 | 0.0000000 | 0.0000000 | 0.4026624 | 1 | 0 |
| rt\_meta | Generalized lithology classified as META: Metamorphic rocks. | 1 (present) ; 0 (absent) | 0 = 18520 1 = 1930 | binary | NA | “NA” | NA | Anning and Ator (2017) | 0.09438 | 0.0000000 | 0.0000000 | 0.2923590 | 1 | 0 |
| rt\_plut\_qtz | Generalized lithology classified as PLUT\_QTZ: Igneous, generally felsic, quartz-rich plutonic rocks such as granitoids, granite, granodacite. Lithology of plutonics and volcanics are split out based on quartz content using the IUGS classification presented in Hyndman’s ‘Petrology of Igneous and Metamorphic Rocks’ second edition, 1985, page 33, fields 1-5. | 1 (present) ; 0 (absent) | 0 = 19475 1 = 975 | binary | NA | “NA” | NA | Anning and Ator (2017) | 0.04768 | 0.0000000 | 0.0000000 | 0.2130877 | 1 | 0 |
| PRMS8110Re | average annual recharge to groundwater | NA | NA | numeric | millimeters per year | “NA” | 1981-2010 | Hay (2019) | 164.37 | 19.2601645 | 254.1999226 | 175.8661000 | 9.994490E+00 | 0.000000E+00 |
| LP2 | Lateral position for fourth order streams | NA | NA | numeric | percent times 100 |  | NA | Moore and others (2019) | 5266 | 2616.2500000 | 7944.0000000 | 3005.1600000 | 10000 | 0 |
| LP4 | Lateral position for second order streams | NA | NA | numeric | percent times 100 |  | NA | Moore and others (2019) | 5098 | 2351.0000000 | 7826.7500000 | 3056.1670000 | 10000 | 0 |
| LP6 | Lateral position for sixth order streams | NA | NA | numeric | percent times 100 |  | NA | Moore and others (2019) | 0.01628 | 2464.0000000 | 7895.7500000 | 3045.5700000 | 10000 | 0 |
| D3 | Devonian; stratified sequence, mainly marine, most complete sequences; Upper Devonian | 1 (present) ; 0 (absent) | 0 = 20117 1 = 333 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.01628 | 0.0000000 | 0.0000000 | 0.1265672 | 1 | 0 |
| De | Devonian eugeosynclinal deposits | 1 (present) ; 0 (absent) | 0 = 20251 1 = 199 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.009731 | 0.0000000 | 0.0000000 | 0.0981674 | 1 | 0 |
| DSe | New England; Devonian and Silurian eugeosynclinal deposits | 1 (present) ; 0 (absent) | 0 = 20003 1 = 447 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.02186 | 0.0000000 | 0.0000000 | 0.1462240 | 1 | 0 |
| Kg | Plutonic and intrusive rocks; Cretaceous granitic rocks; Western United States | 1 (present) ; 0 (absent) | 0 = 20301 1 = 149 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.007286 | 0.0000000 | 0.0000000 | 0.0850490 | 1 | 0 |
| O2 | Ordovician; stratified sequence, mainly marine, most complete sequences; Middle Ordovician (Mohawkian), Upper Ordovician included in places; Eastern United States | 1 (present) ; 0 (absent) | 0 = 20116 1 = 334 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.01633 | 0.0000000 | 0.0000000 | 0.1267539 | 1 | 0 |
| Oe | Ordovician, eugeosynclinal | 1 (present) ; 0 (absent) | 0 = 20232 1 = 218 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.01066 | 0.0000000 | 0.0000000 | 0.1026987 | 1 | 0 |
| PP4 | Upper Pennsylvanian; stratified sequence mainly marine, most complete sequences; Virgilian Series (= Monongahela Group); Central interior region, from Appalachians to Kansas and central Texas | 1 (present) ; 0 (absent) | 0 = 20167 1 = 283 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.01384 | 0.0000000 | 0.0000000 | 0.1168238 | 1 | 0 |
| Pzg1 | Plutonic and intrusive rocks; Lower Paleozoic granitic rocks; about 400 - 500 million years. In New England, Highlandcroft and Oliverian Plutonic Series, related to Taconian orogeny; in Piedmont province | 1 (present) ; 0 (absent) | 0 = 20247 1 = 203 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.009927 | 0.0000000 | 0.0000000 | 0.0991393 | 1 | 0 |
| Pzg2 | Plutonic and intrusive rocks; Middle Paleozoic granitic rocks; about 350-400 million years. In New England, New Hampshire Plutonic Series and others, related to Acadian orogeny; in Piedmont province, synmetamorphic; Appalachian region (New England and Piedmont province), rarely in western United States | 1 (present) ; 0 (absent) | 0 = 20096 1 = 354 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.01731 | 0.0000000 | 0.0000000 | 0.1304289 | 1 | 0 |
| Q | Quaternary; stratified sequence, mainly marine, combinations; Western United States; only large areas and thick deposits shown | 1 (present) ; 0 (absent) | 0 = 17478 1 = 2972 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.1453 | 0.0000000 | 0.0000000 | 0.3524419 | 1 | 0 |
| Qv | Quaternary volcanic rocks | 1 (present) ; 0 (absent) | 0 = 20123 1 = 327 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.01599 | 0.0000000 | 0.0000000 | 0.1254404 | 1 | 0 |
| S2 | Silurian; stratified sequence, mainly marine, most complete sequences; Middle Silurian (Niagaran) | 1 (present) ; 0 (absent) | 0 = 20282 1 = 168 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.008215 | 0.0000000 | 0.0000000 | 0.0902666 | 1 | 0 |
| Se | Silurian eugeosynclinal deposits | 1 (present) ; 0 (absent) | 0 = 20062 1 = 388 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.01897 | 0.0000000 | 0.0000000 | 0.1364333 | 1 | 0 |
| Te2 | Tertiary; Eocene; stratified sequence, mainly marine, most complete sequences; Claiborne Group Atlantic and Gulf Coastal Plains | 1 (present) ; 0 (absent) | 0 = 20027 1 = 423 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.02068 | 0.0000000 | 0.0000000 | 0.1423297 | 1 | 0 |
| Tm | Tertiary; stratified sequence, mainly marine, most complete sequences; Miocene, includes Pliocene in Southeastern U.S. | 1 (present) ; 0 (absent) | 0 = 20096 1 = 354 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.01731 | 0.0000000 | 0.0000000 | 0.1304289 | 1 | 0 |
| Tmc | Tertiary; Miocene continental deposits | 1 (present) ; 0 (absent) | 0 = 20158 1 = 292 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.01428 | 0.0000000 | 0.0000000 | 0.0545356 | 1 | 0 |
| Toc | Tertiary; Oligocene continental deposits | 1 (present) ; 0 (absent) | 0 = 20389 1 = 61 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.002983 | 0.0000000 | 0.0000000 | 0.1186404 | 1 | 0 |
| Tp | Tertiary; stratified sequence, mainly marine, most complete sequences; Pliocene | 1 (present) ; 0 (absent) | 0 = 20309 1 = 141 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.006895 | 0.0000000 | 0.0000000 | 0.0827506 | 1 | 0 |
| Tpc | Tertiary; Pliocene continental deposits includes Miocene and Quaternary deposits in places | 1 (present) ; 0 (absent) | 0 = 19033 1 = 1417 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.06929 | 0.0000000 | 0.0000000 | 0.2539545 | 1 | 0 |
| Tpv | Tertiary; Pliocene volcanic rocks | 1 (present) ; 0 (absent) | 0 = 20285 1 = 165 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.008068 | 0.0000000 | 0.0000000 | 0.0894637 | 1 | 0 |
| Tr | Triassic; stratified sequence, mainly marine, most complete sequences; mainly marine but includes continental Dockum Group in Texas and New Mexico, and Newark Group in Appalachian region | 1 (present) ; 0 (absent) | 0 = 20027 1 = 423 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.02068 | 0.0000000 | 0.0000000 | 0.1423297 | 1 | 0 |
| Txc | Tertiary; Paleocene continental deposits | 1 (present) ; 0 (absent) | 0 = 20237 1 = 213 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.01042 | 0.0000000 | 0.0000000 | 0.1015267 | 1 | 0 |
| uK2 | stratified sequence, mainly marine, most complete sequences; Upper Cretaceous Austin and Eagle Ford Groups (=Colorado Group); Atlantic and Gulf Coastal Plains and western interior region | 1 (present) ; 0 (absent) | 0 = 20049 1 = 401 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.01961 | 0.0000000 | 0.0000000 | 0.1386551 | 1 | 0 |
| uK3 | stratified sequence, mainly marine, most complete sequences; Upper Cretaceous Taylor Group (= Montana Group), In Montana plains, separated into uK3a and b at base of Judith River Formation; Atlantic and Gulf Coastal Plains and western interior region | 1 (present) ; 0 (absent) | 0 = 19932 1 = 518 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.02533 | 0.0000000 | 0.0000000 | 0.1571295 | 1 | 0 |
| uK3b | stratified sequence, mainly marine, most complete sequences; Upper Cretaceous Taylor Group (= Montana Group), In Montana plains, separated into uK3a and b at base of Judith River Formation; Atlantic and Gulf Coastal Plains and western interior region | 1 (present) ; 0 (absent) | 0 = 20417 1 = 33 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.001614 | 0.0000000 | 0.0000000 | 0.0401393 | 1 | 0 |
| uK4 | stratified sequence, mainly marine, most complete sequences; Upper Cretaceous Navarro Group (Laramie and other formations); Atlantic and Gulf Coastal Plains and western interior region | 1 (present) ; 0 (absent) | 0 = 20256 1 = 194 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.009487 | 0.0000000 | 0.0000000 | 0.0969382 | 1 | 0 |
| uPz | stratified sequence, mainly marine, combinations; Upper Paleozoic; Western United States | 1 (present) ; 0 (absent) | 0 = 20250 1 = 200 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.00978 | 0.0000000 | 0.0000000 | 0.0984113 | 1 | 0 |
| Wgn | PreCambrian; metamorphic rocks; Orthogneiss and paragneiss | 1 (present) ; 0 (absent) | 0 = 20389 1 = 61 | binary | NA | “NA” | NA | Schruben and others (1997) | 0.002983 | 0.0000000 | 0.0000000 | 0.0545356 | 1 | 0 |
| as\_c | Arsenic concentration in the soil C horizon | NA | NA | numeric | milligrams per kilogram | “NA” | NA | Smith and others (2014) | 6.9432 | 4.2002700 | 8.2973690 | 7.0556900 | 346.519989 | 0.000032 |
| be\_c | Beryllium concentration in the soil C horizon | NA | NA | numeric | milligrams per kilogram | “NA” | NA | Smith and others (2014) | 1.520005 | 1.0903990 | 1.8808442 | 0.6400278 | 6.937769 | 0.000369 |
| bi\_c | Bismuth concentration in the soil C horizon | NA | NA | numeric | milligrams per kilogram | “NA” | NA | Smith and others (2014) | 0.183433 | 0.1222275 | 0.2208410 | 0.1211548 | 5.138909 | 0.000032 |
| inorgc\_c | Inorganic carbon concentration in the soil C horizon | NA | NA | numeric | weight percent | “NA” | NA | Smith and others (2014) | 0.527658 | 0.0021040 | 0.6033572 | 0.9725496 | 9.999900E-02 | 0.000000E+00 |
| mo\_c | Molybdenum concentration in the soil C horizon | NA | NA | numeric | milligrams per kilogram | “NA” | NA | Smith and others (2014) | 1.03669 | 0.5514590 | 1.1742600 | 1.3213120 | 61.774299 | 0.000179 |
| ni\_c | Nickel concentration in the soil C horizon | NA | NA | numeric | milligrams per kilogram | “NA” | NA | Smith and others (2014) | 29.7553 | 11.5861747 | 26.1105000 | 99.2000500 | 2836.270019 | 0.000031 |
| orgc\_c | Phosphorus concentration in the soil C horizon | NA | NA | numeric | weight percent | “NA” | NA | Smith and others (2014) | 0.88280 | 0.2846100 | 0.9617930 | 1.6647120 | 41.720901 | 0.000031 |
| p\_c | Organic carbon concentration in the soil C horizon | NA | NA | numeric | milligrams per kilogram | “NA” | NA | Smith and others (2014) | 487.883 | 272.1010130 | 622.9609980 | 319.0659000 | 8195.280273 | 0.054348 |
| sb\_c | Antimony concentration in the soil C horizon | NA | NA | numeric | milligrams per kilogram | “NA” | NA | Smith and others (2014) | 0.680345 | 0.3864645 | 0.7895045 | 0.7514792 | 15.530799 | 0.000217 |
| se\_c | Selenium concentration in the soil C horizon | NA | NA | numeric | NA | “NA” | NA | Smith and others (2014) | 0.312980 | 0.2035260 | 0.3555420 | 0.2180525 | 6.379799 | 0.000122 |
| uc\_12 | Alluvial sediments, thick, unit thickness >100 feet, geologic age Holocene to Pliocene | 1 (present) ; 0 (absent) | 0 = 17706 1 = 2740 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.134 | 0.0000000 | 0.0000000 | 0.3406730 | 1 | 0 |
| uc\_221 | Coastal zone sediments, mostly medium-grained, unit thickness <100 feet, geologic age Holocene to Pliocene | 1 (present) ; 0 (absent) | 0 = 20386 1 = 60 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.002935 | 0.0000000 | 0.0000000 | 0.0540934 | 1 | 0 |
| uc\_312 | Eolian sediments, mostly loess, thick, thickness >100 feet, geologic age Holocene to Pliocene | 1 (present) ; 0 (absent) | 0 = 20381 1 = 65 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.003179 | 0.0000000 | 0.0000000 | 0.0562952 | 1 | 0 |
| uc\_321 | Eolian sediments, mostly dune sand, thin, unit thickness <100 feet, geologic age Holocene to Pleistocene | 1 (present) ; 0 (absent) | 0 = 20065 1 = 381 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.01863 | 0.0000000 | 0.0000000 | 0.1352335 | 1 | 0 |
| uc\_322 | Eolian sediments, mostly dune sand, thick, thickness >100 feet, geologic age Holocene to Pleistocene | 1 (present) ; 0 (absent) | 0 = 20403 1 = 43 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.002103 | 0.0000000 | 0.0000000 | 0.0458124 | 1 | 0 |
| uc\_331 | Eolian sediments on southern High Plains, unit thickness <100 feet, geologic age Holocene to Pleistocene | 1 (present) ; 0 (absent) | 0 = 19889 1 = 557 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.02724 | 0.0000000 | 0.0000000 | 0.1627932 | 1 | 0 |
| uc\_421 | Glacial till sediments, mostly silty, thin, unit thickness <100 feet, geologic age late Wisconsinan to pre-Illinoian | 1 (present) ; 0 (absent) | 0 = 19324 1 = 1122 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.05488 | 0.0000000 | 0.0000000 | 0.2277441 | 1 | 0 |
| uc\_422 | Glacial till sediments, mostly silty, thick, unit thickness >100 feet, geologic age late Wisconsinan to pre-Illinoian | 1 (present) ; 0 (absent) | 0 = 19454 1 = 992 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.04852 | 0.0000000 | 0.0000000 | 0.2148635 | 1 | 0 |
| uc\_431 | Glacial till sediments, mostly sandy, thin, unit thickness <100 feet, geologic age late Wisconsinan to pre-Illinoian | 1 (present) ; 0 (absent) | 0 = 19299 1 = 1147 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.0561 | 0.0000000 | 0.0000000 | 0.2301184 | 1 | 0 |
| uc\_451 | Glaciofluvial ice-contact sediments, mostly sand and gravel, thin, unit thickness <100 feet, geologic age late Wisconsinan to Illinoian | 1 (present) ; 0 (absent) | 0 = 20194 1 = 252 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.01232 | 0.0000000 | 0.0000000 | 0.1103351 | 1 | 0 |
| uc\_620 | Colluvial and alluvial sediments, discontinuous, or patchy in distribution, geologic age Holocene to Tertiary | 1 (present) ; 0 (absent) | 0 = 20383 1 = 63 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.003081 | 0.0000000 | 0.0000000 | 0.0554251 | 1 | 0 |
| uc\_811 | Proglacial sediments, mostly fine grained, thin, unit thickness <100 feet, geologic age late Wisconsinan to pre-Illinoian | 1 (present) ; 0 (absent) | 0 = 19622 1 = 824 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.0403 | 0.0000000 | 0.0000000 | 0.1966697 | 1 | 0 |
| uc\_812 | Proglacial sediments, mostly fine grained, thick, unit thickness >100 feet, geologic age late Wisconsinan to pre-Illinoian | 1 (present) ; 0 (absent) | 0 = 20407 1 = 39 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.001907 | 0.0000000 | 0.0000000 | 0.0436339 | 1 | 0 |
| uc\_821 | Proglacial sediments, mostly coarse-grained, thin, unit thickness <100 feet, geologic age late Wisconsinan to pre-Illinoian | 1 (present) ; 0 (absent) | 0 = 19837 1 = 609 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.02979 | 0.0000000 | 0.0000000 | 0.1700000 | 1 | 0 |
| uc\_822 | Proglacial sediments, mostly coarse-grained, thick, unit thickness >100 feet, geologic age late Wisconsinan to pre-Illinoian | 1 (present) ; 0 (absent) | 0 = 19921 1 = 525 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.02568 | 0.0000000 | 0.0000000 | 0.1581749 | 1 | 0 |
| uc\_910 | Residual materials developed in igneous and metamorphic rocks, unit thickness discontinuous, or patchy in distribution, geologic age Holocene to Tertiary | 1 (present) ; 0 (absent) | 0 = 19535 1 = 911 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.04456 | 0.0000000 | 0.0000000 | 0.2063327 | 1 | 0 |
| uc\_920 | Residual materials developed in sedimentary rocks, discontinuous, unit thickness discontinuous, or patchy in distribution, geologic age Holocene to Tertiary | 1 (present) ; 0 (absent) | 0 = 17966 1 = 2480 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.1213 | 0.0000000 | 0.0000000 | 0.3264779 | 1 | 0 |
| uc\_940 | Residual materials developed in carbonate rocks, discontinuous, unit thickness discontinuous, or patchy in distribution, geologic age Holocene to Tertiary | 1 (present) ; 0 (absent) | 0 = 19637 1 = 809 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.03957 | 0.0000000 | 0.0000000 | 0.1949459 | 1 | 0 |
| uc\_960 | Residual materials developed in bedrock, with alluvial sediments, discontinuous, unit thickness discontinuous, or patchy in distribution, geologic age Holocene to Tertiary | 1 (present) ; 0 (absent) | 0 = 20424 1 = 22 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.001076 | 0.0000000 | 0.0000000 | 0.0327857 | 1 | 0 |
| uc\_970 | Residual materials developed in bedrock, discontinuous, unit thickness discontinuous, or patchy in distribution, geologic age Holocene to Tertiary | 1 (present) ; 0 (absent) | 0 = 19655 1 = 791 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.03869 | 0.0000000 | 0.0000000 | 0.1928533 | 1 | 0 |
| uc\_999 | water | 1 (present) ; 0 (absent) | 0 = 20351 1 = 95 NA = 4 | binary | NA | “NA” | NA | Soller and others (2009) | 0.004646 | 0.0000000 | 0.0000000 | 0.0680075 | 1 | 0 |
| Percent\_Ti | County estimates of U.S. tile drainage | NA | NA | numeric | Percent of county | “NA” | 1992 | Sugg (2007) | 1.631 | 0.0000000 | 0.0000000 | 7.7308630 | 70.000000 | 0 |
| DMppt8110 | average annual precipitation from 1981 to 2010 | NA | NA | numeric | inches | “NA” | 1981-2010 | Thornton and others (2018) | 848.37 | 494.2917500 | 1200.3475000 | 387.9446000 | 3098.4300 | 65.2333 |
| na\_10.1 | Cold deserts ecoregion | 1 (present) ; 0 (absent) | 0 = 18260 1 = 2190 | binary | NA | “NA” | NA | USEPA (2013) | 0.1071 | 0.0000000 | 0.0000000 | 0.3092358 | 1 | 0 |
| na\_11.1 | Mediterranean California ecoregion | 1 (present) ; 0 (absent) | 0 = 19284 1 = 1166 | binary | NA | “NA” | NA | USEPA (2013) | 0.05702 | 0.0000000 | 0.0000000 | 0.2318810 | 1 | 0 |
| na\_12.1 | Western Sierra Madre piedmont ecoregion | 1 (present) ; 0 (absent) | 0 = 20366 1 = 84 | binary | NA | “NA” | NA | USEPA (2013) | 0.004108 | 0.0000000 | 0.0000000 | 0.0639602 | 1 | 0 |
| na\_13.1 | Upper Gila Mountains ecoregion | 1 (present) ; 0 (absent) | 0 = 20135 1 = 315 | binary | NA | “NA” | NA | USEPA (2013) | 0.0154 | 0.0000000 | 0.0000000 | 0.1231540 | 1 | 0 |
| na\_6.2 | Western Cordillera ecoregion | 1 (present) ; 0 (absent) | 0 = 18891 1 = 1559 | binary | NA | “NA” | NA | USEPA (2013) | 0.07623 | 0.0000000 | 0.0000000 | 0.2653798 | 1 | 0 |
| na\_7.1 | Marine West Coast forest ecoregion | 1 (present) ; 0 (absent) | 0 = 19776 1 = 674 | binary | NA | “NA” | NA | USEPA (2013) | 0.03296 | 0.0000000 | 0.0000000 | 0.1785322 | 1 | 0 |
| na\_8.1 | Mixed wood plains ecoregion | 1 (present) ; 0 (absent) | 0 = 17535 1 = 2915 | binary | NA | “NA” | NA | USEPA (2013) | 0.1425 | 0.0000000 | 0.0000000 | 0.3496145 | 1 | 0 |
| na\_8.2 | Central USA plains ecoregion | 1 (present) ; 0 (absent) | 0 = 19960 1 = 490 | binary | NA | “NA” | NA | USEPA (2013) | 0.02396 | 0.0000000 | 0.0000000 | 0.1529310 | 1 | 0 |
| na\_9.2 | Temperate prairies ecoregion | 1 (present) ; 0 (absent) | 0 = 19278 1 = 1172 | binary | NA | “NA” | NA | USEPA (2013) | 0.05731 | 0.0000000 | 0.0000000 | 0.2324407 | 1 | 0 |
| na\_9.3 | West-Central semi-arid prairies ecoregion | 1 (present) ; 0 (absent) | 0 = 19941 1 = 509 | binary | NA | “NA” | NA | USEPA (2013) | 0.02489 | 0.0000000 | 0.0000000 | 0.1557936 | 1 | 0 |
| na\_9.4 | South Central semi-arid prairies ecoregion | 1 (present) ; 0 (absent) | 0 = 17115 1 = 3335 | binary | NA | “NA” | NA | USEPA (2013) | 0.1631 | 0.0000000 | 0.0000000 | 0.3694483 | 1 | 0 |
| na\_9.5 | Texas-Louisiana coastal plain ecoregion | 1 (present) ; 0 (absent) | 0 = 20152 1 = 298 | binary | NA | “NA” | NA | USEPA (2013) | 0.01457 | 0.0000000 | 0.0000000 | 0.1198352 | 1 | 0 |
| na\_9.6 | Tamaulipas-Texas semiarid plain ecoregion | 1 (present) ; 0 (absent) | 0 = 20128 1 = 322 | binary | NA | “NA” | NA | USEPA (2013) | 0.01575 | 0.0000000 | 0.0000000 | 0.1244932 | 1 | 0 |
| BFI | Mean annual base flow index | NA | NA | numeric | percent | “NA” | 1950-1981 | Wolock (2003) | 43.56 | 29.0000000 | 57.0000000 | 18.5439400 | 89 | 0 |
| As3Cat | Code to assign observed arsenic concentration to categorical variable | C1 (Arsenic concentration is less than or equal to 5 micrograms per liter); C2 (Arsenic concentration is greater than 5 micrograms per liter and less than or equal to 10 micrograms per liter); C3 (Arsenic concentration is greater than or equal to 10 micrograms per liter) | C1 = 15659 C2 = 1663 C3 = 2383 NA = 745 | categorical | NA | “NA” | NA | Producer defined | (Character variable) | NA | NA | NA | C3 | C1 |
| bas1 | Code to assign observed arsenic concentration to binary variable | 1 (Arsenic concentration is greater than 1 microgram per liter); 0 (Arsenic concentration is less than or equal to 1 microgram per liter) | 0 = 8750 1 = 8108 NA = 3592 | binary | NA | “NA” | NA | Producer defined | 0.481 | 0.0000000 | 1.0000000 | 0.4996521 | 1 | 0 |
| bas10 | Code to assign observed arsenic concentration to binary variable | 1 (Arsenic concentration is greater than 10 micrograms per liter); 0 (Arsenic concentration is less than or equal to 10 micrograms per liter) | 0 = 18241 1 = 2209 | binary | NA | “NA” | NA | Producer defined | 0.108 | 0.0000000 | 0.0000000 | 0.3104127 | 1 | 0 |
| bas5 | Code to assign observed arsenic concentration to binary variable | 1 (Arsenic concentration is greater than 5 micrograms per liter); 0 (Arsenic concentration is less than or equal to 5 micrograms per liter) | 0 = 15874 1 = 3831 NA = 745 | binary | NA | “NA” | NA | Producer defined | 0.1944 | 0.0000000 | 0.0000000 | 0.3957618 | 1 | 0 |
| spl1 | Code to indicate if the location was used in the boosted regression tree As 1 model training or testing dataset | TRUE (Used in model training dataset for the boosted regression trees model predicting arsenic greater than 1 microgram per liter); FALSE (Used in model testing dataset for the boosted regression trees model predicting arsenic greater than 1 microgram per liter) | FALSE = 5057 TRUE = 11801 NA = 3592 | categorical | NA | “NA” | NA | Producer defined | (Logical variable) | NA | NA | NA | TRUE | FALSE |
| spl10 | Code to indicate if the location was used in the boosted regression tree As 10 model training or testing dataset | TRUE (Used in model training dataset for the boosted regression trees model predicting arsenic greater than 10 micrograms per liter); FALSE (Used in model testing dataset for the boosted regression trees model predicting arsenic greater than 10 micrograms per liter) | FALSE = 6135 TRUE = 14315 | categorical | NA | “NA” | NA | Producer defined | (Logical variable) | NA | NA | NA | TRUE | FALSE |
| spl3cat | Code to indicate if the location was used in the random forest model training or testing dataset | TRUE (Used in model training dataset for the random forest classification model); FALSE (Used in model testing dataset for the random forest classification model) | FALSE = 5912 TRUE = 13793 NA = 745 | categorical | NA | “NA” | NA | Producer defined | (Logical variable) | NA | NA | NA | TRUE | FALSE |
| spl5 | Code to indicate if the location was used in the boosted regression tree As 5 model training or testing dataset | TRUE (Used in model training dataset for the boosted regression trees model predicting arsenic greater than 5 micrograms per liter); FALSE (Used in model testing dataset for the boosted regression trees model predicting arsenic greater than 5 micrograms per liter) | FALSE = 5911 TRUE = 13794 NA = 745 | categorical | NA | “NA” | NA | Producer defined | (Logical variable) | NA | NA | NA | TRUE | FALSE |

# Categorical Variables stats summary

# library(kableExtra) #only need to load the library once  
#setwd("/Users/austinmartinez/Documents/GitHub/coPlateauWaterQuality/01\_data")  
setwd("C:/Users/jhoover/Documents/GitHub/coPlateauWaterQuality/01\_data")  
arsenic\_catagorical\_data <- read.csv("Arsenic\_catagorical\_variables.csv")  
# Extracting specific columns  
catagorical\_data <- arsenic\_catagorical\_data[, c("Predictor\_name", "Category\_levels", "Categorical\_observation\_amounts")]  
kable(catagorical\_data)

| Predictor\_name | Category\_levels | Categorical\_observation\_amounts |
| --- | --- | --- |
| rt\_carb | 1 (present) ; 0 (absent) | 0 = 17306 1 = 3144 |
| rt\_clast\_c | 1 (present) ; 0 (absent) | 0 = 15343 1 = 5107 |
| rt\_clast\_f | 1 (present) ; 0 (absent) | 0 = 16422 1 = 4028 |
| rt\_clast\_u | 1 (present) ; 0 (absent) | 0 = 16287 1 = 4163 |
| rt\_meta | 1 (present) ; 0 (absent) | 0 = 18520 1 = 1930 |
| rt\_plut\_qtz | 1 (present) ; 0 (absent) | 0 = 19475 1 = 975 |
| D3 | 1 (present) ; 0 (absent) | 0 = 20117 1 = 333 |
| De | 1 (present) ; 0 (absent) | 0 = 20251 1 = 199 |
| DSe | 1 (present) ; 0 (absent) | 0 = 20003 1 = 447 |
| Kg | 1 (present) ; 0 (absent) | 0 = 20301 1 = 149 |
| O2 | 1 (present) ; 0 (absent) | 0 = 20116 1 = 334 |
| Oe | 1 (present) ; 0 (absent) | 0 = 20232 1 = 218 |
| PP4 | 1 (present) ; 0 (absent) | 0 = 20167 1 = 283 |
| Pzg1 | 1 (present) ; 0 (absent) | 0 = 20247 1 = 203 |
| Pzg2 | 1 (present) ; 0 (absent) | 0 = 20096 1 = 354 |
| Q | 1 (present) ; 0 (absent) | 0 = 17478 1 = 2972 |
| Qv | 1 (present) ; 0 (absent) | 0 = 20123 1 = 327 |
| S2 | 1 (present) ; 0 (absent) | 0 = 20282 1 = 168 |
| Se | 1 (present) ; 0 (absent) | 0 = 20062 1 = 388 |
| Te2 | 1 (present) ; 0 (absent) | 0 = 20027 1 = 423 |
| Tm | 1 (present) ; 0 (absent) | 0 = 20096 1 = 354 |
| Tmc | 1 (present) ; 0 (absent) | 0 = 20158 1 = 292 |
| Toc | 1 (present) ; 0 (absent) | 0 = 20389 1 = 61 |
| Tp | 1 (present) ; 0 (absent) | 0 = 20309 1 = 141 |
| Tpc | 1 (present) ; 0 (absent) | 0 = 19033 1 = 1417 |
| Tpv | 1 (present) ; 0 (absent) | 0 = 20285 1 = 165 |
| Tr | 1 (present) ; 0 (absent) | 0 = 20027 1 = 423 |
| Txc | 1 (present) ; 0 (absent) | 0 = 20237 1 = 213 |
| uK2 | 1 (present) ; 0 (absent) | 0 = 20049 1 = 401 |
| uK3 | 1 (present) ; 0 (absent) | 0 = 19932 1 = 518 |
| uK3b | 1 (present) ; 0 (absent) | 0 = 20417 1 = 33 |
| uK4 | 1 (present) ; 0 (absent) | 0 = 20256 1 = 194 |
| uPz | 1 (present) ; 0 (absent) | 0 = 20250 1 = 200 |
| Wgn | 1 (present) ; 0 (absent) | 0 = 20389 1 = 61 |
| uc\_12 | 1 (present) ; 0 (absent) | 0 = 17706 1 = 2740 NA = 4 |
| uc\_221 | 1 (present) ; 0 (absent) | 0 = 20386 1 = 60 NA = 4 |
| uc\_312 | 1 (present) ; 0 (absent) | 0 = 20381 1 = 65 NA = 4 |
| uc\_321 | 1 (present) ; 0 (absent) | 0 = 20065 1 = 381 NA = 4 |
| uc\_322 | 1 (present) ; 0 (absent) | 0 = 20403 1 = 43 NA = 4 |
| uc\_331 | 1 (present) ; 0 (absent) | 0 = 19889 1 = 557 NA = 4 |
| uc\_421 | 1 (present) ; 0 (absent) | 0 = 19324 1 = 1122 NA = 4 |
| uc\_422 | 1 (present) ; 0 (absent) | 0 = 19454 1 = 992 NA = 4 |
| uc\_431 | 1 (present) ; 0 (absent) | 0 = 19299 1 = 1147 NA = 4 |
| uc\_451 | 1 (present) ; 0 (absent) | 0 = 20194 1 = 252 NA = 4 |
| uc\_620 | 1 (present) ; 0 (absent) | 0 = 20383 1 = 63 NA = 4 |
| uc\_811 | 1 (present) ; 0 (absent) | 0 = 19622 1 = 824 NA = 4 |
| uc\_812 | 1 (present) ; 0 (absent) | 0 = 20407 1 = 39 NA = 4 |
| uc\_821 | 1 (present) ; 0 (absent) | 0 = 19837 1 = 609 NA = 4 |
| uc\_822 | 1 (present) ; 0 (absent) | 0 = 19921 1 = 525 NA = 4 |
| uc\_910 | 1 (present) ; 0 (absent) | 0 = 19535 1 = 911 NA = 4 |
| uc\_920 | 1 (present) ; 0 (absent) | 0 = 17966 1 = 2480 NA = 4 |
| uc\_940 | 1 (present) ; 0 (absent) | 0 = 19637 1 = 809 NA = 4 |
| uc\_960 | 1 (present) ; 0 (absent) | 0 = 20424 1 = 22 NA = 4 |
| uc\_970 | 1 (present) ; 0 (absent) | 0 = 19655 1 = 791 NA = 4 |
| uc\_999 | 1 (present) ; 0 (absent) | 0 = 20351 1 = 95 NA = 4 |
| na\_10.1 | 1 (present) ; 0 (absent) | 0 = 18260 1 = 2190 |
| na\_11.1 | 1 (present) ; 0 (absent) | 0 = 19284 1 = 1166 |
| na\_12.1 | 1 (present) ; 0 (absent) | 0 = 20366 1 = 84 |
| na\_13.1 | 1 (present) ; 0 (absent) | 0 = 20135 1 = 315 |
| na\_6.2 | 1 (present) ; 0 (absent) | 0 = 18891 1 = 1559 |
| na\_7.1 | 1 (present) ; 0 (absent) | 0 = 19776 1 = 674 |
| na\_8.1 | 1 (present) ; 0 (absent) | 0 = 17535 1 = 2915 |
| na\_8.2 | 1 (present) ; 0 (absent) | 0 = 19960 1 = 490 |
| na\_9.2 | 1 (present) ; 0 (absent) | 0 = 19278 1 = 1172 |
| na\_9.3 | 1 (present) ; 0 (absent) | 0 = 19941 1 = 509 |
| na\_9.4 | 1 (present) ; 0 (absent) | 0 = 17115 1 = 3335 |
| na\_9.5 | 1 (present) ; 0 (absent) | 0 = 20152 1 = 298 |
| na\_9.6 | 1 (present) ; 0 (absent) | 0 = 20128 1 = 322 |
| As3Cat | C1 (Arsenic concentration is less than or equal to 5 micrograms per liter); C2 (Arsenic concentration is greater than 5 micrograms per liter and less than or equal to 10 micrograms per liter); C3 (Arsenic concentration is greater than or equal to 10 micrograms per liter) | C1 = 15659 C2 = 1663 C3 = 2383 NA = 745 |
| bas1 | 1 (Arsenic concentration is greater than 1 microgram per liter); 0 (Arsenic concentration is less than or equal to 1 microgram per liter) | 0 = 8750 1 = 8108 NA = 3592 |
| bas10 | 1 (Arsenic concentration is greater than 10 micrograms per liter); 0 (Arsenic concentration is less than or equal to 10 micrograms per liter) | 0 = 18241 1 = 2209 |
| bas5 | 1 (Arsenic concentration is greater than 5 micrograms per liter); 0 (Arsenic concentration is less than or equal to 5 micrograms per liter) | 0 = 15874 1 = 3831 NA = 745 |
| spl1 | TRUE (Used in model training dataset for the boosted regression trees model predicting arsenic greater than 1 microgram per liter); FALSE (Used in model testing dataset for the boosted regression trees model predicting arsenic greater than 1 microgram per liter) | FALSE = 5057 TRUE = 11801 NA = 3592 |
| spl10 | TRUE (Used in model training dataset for the boosted regression trees model predicting arsenic greater than 10 micrograms per liter); FALSE (Used in model testing dataset for the boosted regression trees model predicting arsenic greater than 10 micrograms per liter) | FALSE = 6135 TRUE = 14315 |
| spl3cat | TRUE (Used in model training dataset for the random forest classification model); FALSE (Used in model testing dataset for the random forest classification model) | FALSE = 5912 TRUE = 13793 NA = 745 |
| spl5 | TRUE (Used in model training dataset for the boosted regression trees model predicting arsenic greater than 5 micrograms per liter); FALSE (Used in model testing dataset for the boosted regression trees model predicting arsenic greater than 5 micrograms per liter) | FALSE = 5911 TRUE = 13794 NA = 745 |

# Continuous Variables stats summary

#library(kableExtra)  
#setwd("/Users/austinmartinez/Documents/GitHub/coPlateauWaterQuality/01\_data")  
setwd("C:/Users/jhoover/Documents/GitHub/coPlateauWaterQuality/01\_data")  
arsenic\_continous\_data <- read.csv("Arsenic\_continuos\_variables.csv")  
# Extracting specific columns  
extracted\_data <- arsenic\_continous\_data[, c("Predictor\_name", "Min", "Max", "X25.\_Quartile", "Mean", "X75.\_Quartile",  
 "Standard\_Deviation")]  
kable(extracted\_data)

| Predictor\_name | Min | Max | X25.\_Quartile | Mean | X75.\_Quartile | Standard\_Deviation |
| --- | --- | --- | --- | --- | --- | --- |
| PRMS8110Re | 0.000000 | 9.994490 | 19.2601645 | 164.370000 | 254.1999226 | 175.8661000 |
| LP2 | 0.000000 | 10000.000000 | 2616.2500000 | 5266.000000 | 7944.0000000 | 3005.1600000 |
| LP4 | 0.000000 | 10000.000000 | 2351.0000000 | 5098.000000 | 7826.7500000 | 3056.1670000 |
| LP6 | 0.000000 | 10000.000000 | 2464.0000000 | 0.016280 | 7895.7500000 | 3045.5700000 |
| as\_c | 0.000032 | 346.519989 | 4.2002700 | 6.943200 | 8.2973690 | 7.0556900 |
| be\_c | 0.000369 | 6.937769 | 1.0903990 | 1.520005 | 1.8808442 | 0.6400278 |
| bi\_c | 0.000032 | 5.138909 | 0.1222275 | 0.183433 | 0.2208410 | 0.1211548 |
| inorgc\_c | 0.000000 | 0.099999 | 0.0021040 | 0.527658 | 0.6033572 | 0.9725496 |
| mo\_c | 0.000179 | 61.774299 | 0.5514590 | 1.036690 | 1.1742600 | 1.3213120 |
| ni\_c | 0.000031 | 2836.270019 | 11.5861747 | 29.755300 | 26.1105000 | 99.2000500 |
| orgc\_c | 0.000031 | 41.720901 | 0.2846100 | 0.882800 | 0.9617930 | 1.6647120 |
| p\_c | 0.054348 | 8195.280273 | 272.1010130 | 487.883000 | 622.9609980 | 319.0659000 |
| sb\_c | 0.000217 | 15.530799 | 0.3864645 | 0.680345 | 0.7895045 | 0.7514792 |
| se\_c | 0.000122 | 6.379799 | 0.2035260 | 0.312980 | 0.3555420 | 0.2180525 |
| Percent\_Ti | 0.000000 | 70.000000 | 0.0000000 | 1.631000 | 0.0000000 | 7.7308630 |
| DMppt8110 | 65.233300 | 3098.430000 | 494.2917500 | 848.370000 | 1200.3475000 | 387.9446000 |
| BFI | 0.000000 | 89.000000 | 29.0000000 | 43.560000 | 57.0000000 | 18.5439400 |