Water Quality Session

Kateri Salk

2021-09-22

Intro

Water quality monitoring data can be used across multiple facets of water management, including exploration, stressor-response analysis, assessment, and TMDL development. One of the most comprehensive repositories for water quality data is the Water Quality Portal. The Water Quality Portal contains data collected by over 400 state, federal, tribal, and local agencies, including EPA STORET data and USGS NWIS data. Processing and analyzing NOAA data is an ideal process to automate, since the data downloaded from NOAA's platform have a consistent format.

During this session, we will:

1. Import Water Quality Portal data into R
2. Automate common processing and quality assurance steps
3. Wrangle water quality data
4. Visualize processed water quality data

Setup

Acquiring Water Quality Portal Data

the `dataRetrieval` package not only allows us to gather hydrologic information from USGS gage sites, but also water quality data from the Water Quality Portal. We will be using just two of the functions for this session, but there are several great resources that outline the wide-ranging functionality of the package listed below.

Introduction to the dataRetrieval package General Data Import from Water Quality Portal Water Quality Portal Web Services Guide dataRetrieval Tutorial

```
ManitowocWQ <- readWQPdata(siteid = c("WIDNR_WQX-363219", "WIDNR_WQX-363069"))
ManitowocSites <- whatWQPsites(siteid = c("WIDNR_WQX-363219", "WIDNR_WQX-363069"))
```

Data Processing

Site Metadata

`select` allows us to subset columns of a dataset. Use a colon to specify a range of columns, and commas to specify individual columns.

```
ManitowocSites_subset <- ManitowocSites %>%
  select(OrganizationIdentifier:MonitoringLocationName, MonitoringLocationDescriptionText,
         HUCEightDigitCode, LatitudeMeasure, LongitudeMeasure)
```

Water Quality data

Water Quality Portal downloads have the same columns each time, but be aware that data are uploaded to the Water Quality Portal by individual organizations, which may or may not follow the same conventions. Data

and metadata quality are not guaranteed! Make sure to carefully explore any data and make conservative quality assurance decisions where information is limited.

General data processing and quality assurance considerations:

1. WQP data is acquired in long format. It may be useful to wrangle the dataset into wide format (we will do this today)
2. `readWQPdata` does not inherently restrict the variables pulled from WQP. You may specify the desired variables by using, for instance: 'characteristicName = "pH" '
3. **ResultMeasureValue** should be numeric, with details on detection limits, qualifiers, etc. provided in other columns. This is not always the case!
4. **ResultSampleFractionText** specifies forms of constituents. In some cases, a single **CharacteristicName** will have both "Total" and "Dissolved" forms specified, which should not be combined.
5. Some variables have different names but represent the same constituent (e.g., "Total Kjeldahl nitrogen (Organic N & NH3)" and "Kjeldahl nitrogen"). Always refer to the **ResultAnalyticalMethod** columns to verify methods are measuring the same constituent.
6. **ActivityDepthHeightMeasure.MeasureValue** provides depth information. This is a crucial column for lake data but less often for river data.
7. **ResultCommentText** often has details relating to additional QA.
8. **MeasureQualifierCode** Contains information about data flags:

- *U* designates below detection limit (action: set value to 1/2 detection or quantitation limit from **DetectionQuantitationLimitMeasure.MeasureValue**)
- *J* designates above detection limit but below quantitation limit (action: retain value)
- Other codes may designate suspect data or other flags which may be described in detail in **ResultLaboratoryCommentText** or another column

Wrangling functions we will use (feel free to add notes here or comment in the code):

- `filter`
- `mutate`
- `select`
- `group_by`
- `summarise`
- `left_join`

```
View(ManitowocWQ)

ManitowocWQ$ActivityStartDate <- as.Date(ManitowocWQ$ActivityStartDate, format = "%Y-%m-%d")
unique(ManitowocWQ$CharacteristicName)
```

```
##  [1] "Temperature, air"
##  [2] "Escherichia coli"
##  [3] "Cloud cover"
##  [4] "Dissolved oxygen (DO)"
##  [5] "Inorganic nitrogen (nitrate and nitrite)"
##  [6] "Orthophosphate"
##  [7] "Phosphate-phosphorus"
##  [8] "Specific conductance"
##  [9] "Total suspended solids"
## [10] "pH"
## [11] "Kjeldahl nitrogen"
## [12] "Nitrate + Nitrite"
## [13] "Phosphorus"
## [14] "Total Kjeldahl nitrogen (Organic N & NH3)"
## [15] "Calcium carbonate"
## [16] "Chloride"
```

```
## [17] "Dissolved oxygen saturation"
## [18] "Temperature, water"
## [19] "Chlorophyll a (probe relative fluorescence)"
## [20] "Ammonia"
## [21] "Turbidity"
## [22] "Silica"
## [23] "Transparency, tube with disk"
## [24] "Suspended Sediment Concentration (SSC)"
## [25] "Biochemical oxygen demand, standard conditions"
## [26] "Total fixed solids"
## [27] "Fecal Coliform"
## [28] "Sodium"
## [29] "Fecal Streptococcus Group Bacteria"
## [30] "Potassium"
## [31] "Calcium"
## [32] "True color"
## [33] "Chemical oxygen demand"
## [34] "Chlorophyll a, uncorrected for pheophytin"
## [35] "Sulfate"
## [36] "Flow"
## [37] "Magnesium"
## [38] "Hardness, Ca, Mg"
## [39] "Count"
```

```
# Some cells in ResultMeasureValue have * or ND noted.
# Since no columns are available to tell us what these codes mean, we will set these values to NA.
# Setting the column to numeric will set any cells containing non-numeric characters to NA.
class(ManitowocWQ$ResultMeasureValue)
```

```
## [1] "character"
```

```
ManitowocWQ$ResultMeasureValue <- as.numeric(ManitowocWQ$ResultMeasureValue)
```

```
## Warning: NAs introduced by coercion
```

```
# example: if MeasureQualifierCode has a "U", set value to 1/2 quantitation limit
# example 1:
# ManitowocWQ <- ManitowocWQ %>%
#   mutate(ResultMeasureValue = case_when(MeasureQualifierCode == "U" ~ DetectionQuantitationLimitMeasu
#                                         TRUE ~ ResultMeasureValue))
# example 2:
# ManitowocWQ$ResultMeasureValue[ManitowocWQ$MeasureQualifierCode == "U"] <-
#   ManitowocWQ$DetectionQuantitationLimitMeasure.MeasureValue/2

# example: generate a dataset for only one constituent
ManitowocWQ_pH <- ManitowocWQ %>%
  filter(CharacteristicName == "pH")


ManitowocWQ_long <- ManitowocWQ %>%
  # filter pH, suspended solids, DO, nutrients, and chlorophyll
  filter(CharacteristicName %in% c("pH", "Total suspended solids", "Turbidity",
                                   "Suspended Sediment Concentration (SSC)",
                                   "Dissolved oxygen (DO)", "Dissolved oxygen saturation",
                                   "Kjeldahl nitrogen", "Ammonia", "Nitrate + Nitrite",
                                   "Inorganic nitrogen (nitrate and nitrite)",
```

```
                                    "Total Kjeldahl nitrogen (Organic N & NH3)",
                                    "Orthophosphate", "Phosphate-phosphorus", "Phosphorus",
                                    "Chlorophyll a (probe relative fluorescence)",
                                    "Chlorophyll a, uncorrected for pheophytin")) %>%
  # re-name variables with no spaces, assign differently named variables as the same
  # add units in the name. Units are typically provided in ResultMeasure.MeasureUnitCode
  mutate(Variable = case_when(CharacteristicName == "pH" ~ "pH",
                              CharacteristicName == "Total suspended solids" ~ "TSS_mgL",
                              CharacteristicName == "Suspended Sediment Concentration (SSC)" ~ "TSS_mgL",
                              CharacteristicName == "Dissolved oxygen (DO)" ~ "DO_mgL",
                              CharacteristicName == "Dissolved oxygen saturation" ~ "DO_mgL",
                              CharacteristicName == "Kjeldahl nitrogen" ~ "TKN_mgL",
                              CharacteristicName == "Ammonia" &  ResultSampleFractionText == "Dissolved",
                              CharacteristicName == "Ammonia" &  ResultSampleFractionText == "Total" ~ a
                              CharacteristicName == "Nitrate + Nitrite" ~ "NO23_mgL",
                              CharacteristicName == "Inorganic nitrogen (nitrate and nitrite)" ~ "NO23_m
                              CharacteristicName == "Total Kjeldahl nitrogen (Organic N & NH3)" ~ "TKN_m
                              CharacteristicName == "Orthophosphate" ~ "Orthophosphate_mgL",
                              CharacteristicName == "Phosphate-phosphorus" &  ResultSampleFractionText =
                              CharacteristicName == "Phosphate-phosphorus" &  ResultSampleFractionText =
                              CharacteristicName == "Chlorophyll a (probe relative fluorescence)" ~ "Chl
                              CharacteristicName == "Chlorophyll a, uncorrected for pheophytin" ~ "Chla_
  select(OrganizationIdentifier, OrganizationFormalName, ActivityStartDate,
         ActivityConductingOrganizationText, MonitoringLocationIdentifier,
         ActivityDepthHeightMeasure.MeasureValue, ResultMeasureValue, Variable) %>%
  group_by(OrganizationIdentifier, OrganizationFormalName, ActivityStartDate,
         ActivityConductingOrganizationText, MonitoringLocationIdentifier,
         ActivityDepthHeightMeasure.MeasureValue, Variable) %>%
  summarise(ResultMeasureValue = mean(ResultMeasureValue, na.rm = TRUE)) %>%
  mutate(Month = month(ActivityStartDate),
         Year = year(ActivityStartDate))
```

```
## `summarise()` has grouped output by 'OrganizationIdentifier', 'OrganizationFormalName', 'ActivitySta
```

```
ManitowocWQ_wide <- ManitowocWQ_long %>%
  pivot_wider(names_from = "Variable", values_from = "ResultMeasureValue") %>%
  filter(ActivityDepthHeightMeasure.MeasureValue <= 1)
```

Join data and metadata

```
ManitowocWQ_wide <- left_join(ManitowocWQ_wide, ManitowocSites_subset)
```

```
## Joining, by = c("OrganizationIdentifier", "OrganizationFormalName", "MonitoringLocationIdentifier")
```

Exploratory data analysis

```
str(ManitowocWQ_wide)
```

```
## grouped_df [287 x 25] (S3: grouped_df/tbl_df/tbl/data.frame)
##  $ OrganizationIdentifier                : chr [1:287] "WIDNR_WQX" "WIDNR_WQX" "WIDNR_WQX" "WIDNR_WQ
##  $ OrganizationFormalName                : chr [1:287] "Wisconsin Department of Natural Resources"
##  $ ActivityStartDate                     : Date[1:287], format: "1996-06-20" "1997-01-27" ...
##  $ ActivityConductingOrganizationText    : chr [1:287] "WIDNR_WQX" "WIDNR_WQX" "WIDNR_WQX" "WIDNR_WQ
##  $ MonitoringLocationIdentifier          : chr [1:287] "WIDNR_WQX-363069" "WIDNR_WQX-363069" "WIDNR_
```

```
##  $ ActivityDepthHeightMeasure.MeasureValue: num [1:287] 1 1 1 1 1 1 1 1 1 1 ...
##  $ Month                                  : num [1:287] 6 1 4 9 4 6 7 11 1 4 ...
##  $ Year                                   : num [1:287] 1996 1997 1997 1997 1998 ...
##  $ Chla_uncorrected_ugL                   : num [1:287] NA NA NA NA NA NA NA NA NA NA ...
##  $ DO_mgL                                 : num [1:287] 3.51 6.3 14 8.2 10.9 9.95 8.1 12.2 NA NA ...
##  $ NO23_mgL                               : num [1:287] 1.8 2.57 0.555 0.134 2.41 0.786 0.293 1.32 4
##  $ Orthophosphate_mgL                     : num [1:287] 0.165 0.117 0.029 NaN 0.145 0.075 0.053 0.003
##  $ pH                                     : num [1:287] 7.75 8.1 8.34 NA NA ...
##  $ TKN_mgL                                : num [1:287] NA 1.5 1.8 1.3 NaN NaN 1.17 1.57 1.16 1.32 .
##  $ TP_mgL                                 : num [1:287] NaN 0.153 0.156 0.179 0.387 0.228 0.134 0.155
##  $ NA                                     : num [1:287] 0.193 NA NA NA NA NA NA NA 2 13.5 ...
##  $ NH3_mgL                                : num [1:287] NA 0.414 NaN 0.029 NaN 0.2 0.046 NaN 0.115 0
##  $ Chl_probe_RFU                          : num [1:287] NA NA NA NA NA NA NA NA NA NA ...
##  $ TSS_mgL                                : num [1:287] NA NA NA NA NA NA NA NA NA NA ...
##  $ TDP_mgL                                : num [1:287] NA NA NA NA NA NA NA NA NA NA ...
##  $ MonitoringLocationName                 : chr [1:287] "Manitowoc River at Cth Jj(Michigan Ave)" "Ma
##  $ MonitoringLocationDescriptionText      : chr [1:287] "AT USGS GAGING STATION 04085427. MONTHLY MOI
##  $ HUCEightDigitCode                      : chr [1:287] "04030101" "04030101" "04030101" "04030101"
##  $ LatitudeMeasure                        : num [1:287] 44.1 44.1 44.1 44.1 44.1 ...
##  $ LongitudeMeasure                       : num [1:287] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  - attr(*, "groups")= tibble [287 x 7] (S3: tbl_df/tbl/data.frame)
##   ..$ OrganizationIdentifier                : chr [1:287] "WIDNR_WQX" "WIDNR_WQX" "WIDNR_WQX" "WIDNI
##   ..$ OrganizationFormalName                : chr [1:287] "Wisconsin Department of Natural Resources
##   ..$ ActivityStartDate                     : Date[1:287], format: "1996-06-20" "1997-01-27" ...
##   ..$ ActivityConductingOrganizationText    : chr [1:287] "WIDNR_WQX" "WIDNR_WQX" "WIDNR_WQX" "WIDNI
##   ..$ MonitoringLocationIdentifier          : chr [1:287] "WIDNR_WQX-363069" "WIDNR_WQX-363069" "WI
##   ..$ ActivityDepthHeightMeasure.MeasureValue: num [1:287] 1 1 1 1 1 1 1 1 1 1 ...
##   ..$ .rows                                 : list<int> [1:287]
##   .. ..$ : int 1
##   .. ..$ : int 2
##   .. ..$ : int 3
##   .. ..$ : int 4
##   .. ..$ : int 5
##   .. ..$ : int 6
##   .. ..$ : int 7
##   .. ..$ : int 8
##   .. ..$ : int 9
##   .. ..$ : int 10
##   .. ..$ : int 11
##   .. ..$ : int 12
##   .. ..$ : int 13
##   .. ..$ : int 14
##   .. ..$ : int 15
##   .. ..$ : int 16
##   .. ..$ : int 17
##   .. ..$ : int 18
##   .. ..$ : int 19
##   .. ..$ : int 20
##   .. ..$ : int 21
##   .. ..$ : int 22
##   .. ..$ : int 23
##   .. ..$ : int 24
##   .. ..$ : int 25
##   .. ..$ : int 26
```

```
##   .. ..$ : int 27
##   .. ..$ : int 28
##   .. ..$ : int 29
##   .. ..$ : int 30
##   .. ..$ : int 31
##   .. ..$ : int 32
##   .. ..$ : int 33
##   .. ..$ : int 34
##   .. ..$ : int 35
##   .. ..$ : int 36
##   .. ..$ : int 37
##   .. ..$ : int 38
##   .. ..$ : int 39
##   .. ..$ : int 40
##   .. ..$ : int 41
##   .. ..$ : int 42
##   .. ..$ : int 43
##   .. ..$ : int 44
##   .. ..$ : int 45
##   .. ..$ : int 46
##   .. ..$ : int 47
##   .. ..$ : int 48
##   .. ..$ : int 49
##   .. ..$ : int 50
##   .. ..$ : int 51
##   .. ..$ : int 52
##   .. ..$ : int 53
##   .. ..$ : int 54
##   .. ..$ : int 55
##   .. ..$ : int 56
##   .. ..$ : int 57
##   .. ..$ : int 58
##   .. ..$ : int 59
##   .. ..$ : int 60
##   .. ..$ : int 61
##   .. ..$ : int 62
##   .. ..$ : int 63
##   .. ..$ : int 64
##   .. ..$ : int 65
##   .. ..$ : int 66
##   .. ..$ : int 67
##   .. ..$ : int 68
##   .. ..$ : int 69
##   .. ..$ : int 70
##   .. ..$ : int 71
##   .. ..$ : int 72
##   .. ..$ : int 73
##   .. ..$ : int 74
##   .. ..$ : int 75
##   .. ..$ : int 76
##   .. ..$ : int 77
##   .. ..$ : int 78
##   .. ..$ : int 79
##   .. ..$ : int 80
```

```
##   .. ..$ : int 81
##   .. ..$ : int 82
##   .. ..$ : int 83
##   .. ..$ : int 84
##   .. ..$ : int 85
##   .. ..$ : int 86
##   .. ..$ : int 87
##   .. ..$ : int 88
##   .. ..$ : int 89
##   .. ..$ : int 90
##   .. ..$ : int 91
##   .. ..$ : int 92
##   .. ..$ : int 93
##   .. ..$ : int 94
##   .. ..$ : int 95
##   .. ..$ : int 96
##   .. ..$ : int 97
##   .. ..$ : int 98
##   .. ..$ : int 99
##   .. .. [list output truncated]
##   .. ..@ ptype: int(0)
##   ..- attr(*, ".drop")= logi TRUE
```

```
summary(ManitowocWQ_wide)
```

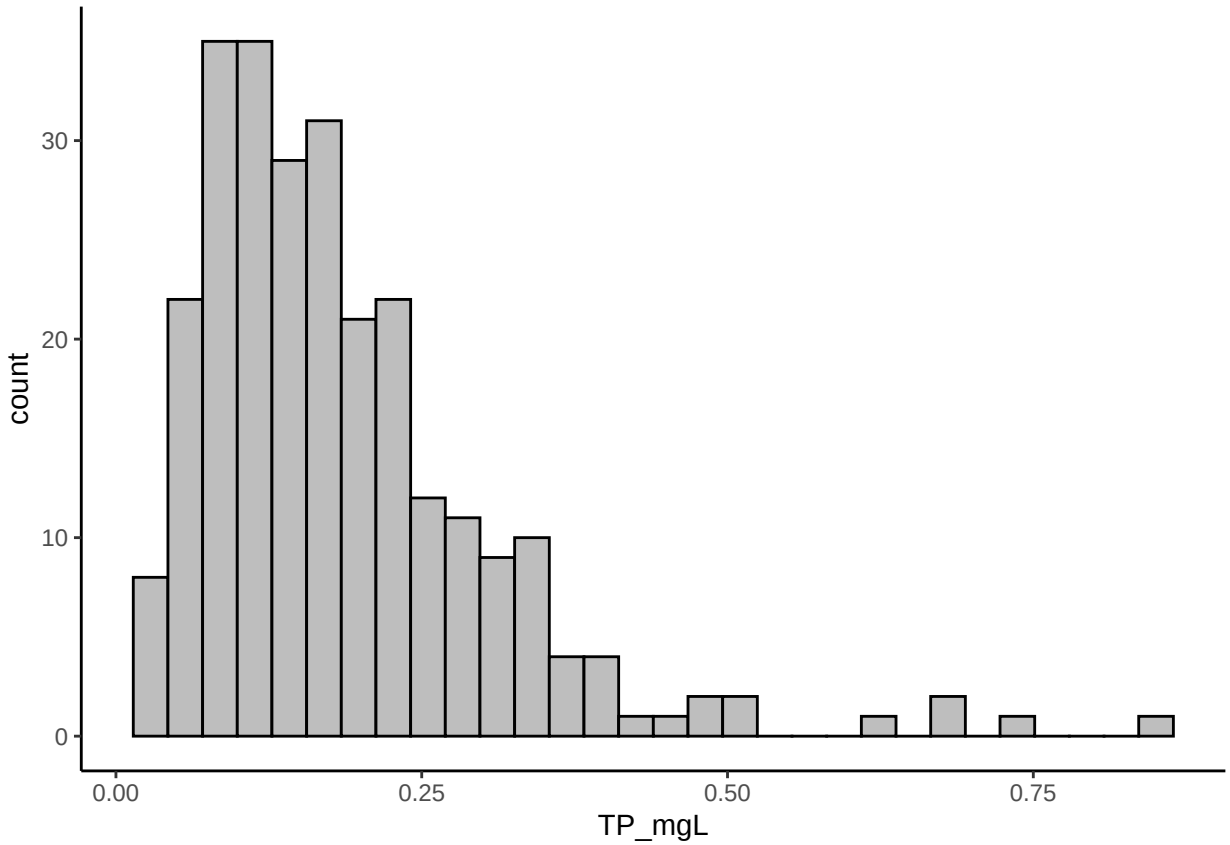
```
##  OrganizationIdentifier OrganizationFormalName ActivityStartDate
##  Length:287             Length:287             Min.   :1996-06-20
##  Class :character       Class :character       1st Qu.:2009-04-19
##  Mode  :character       Mode  :character       Median :2012-10-15
##                                                Mean   :2012-08-18
##                                                3rd Qu.:2016-04-22
##                                                Max.   :2021-07-14
##
##  ActivityConductingOrganizationText MonitoringLocationIdentifier
##  Length:287                         Length:287
##  Class :character                   Class :character
##  Mode  :character                   Mode  :character
##
##
##
##
##  ActivityDepthHeightMeasure.MeasureValue     Month             Year
##  Min.   :0.1000                          Min.   : 1.000   Min.   :1996
##  1st Qu.:0.1000                          1st Qu.: 3.000   1st Qu.:2009
##  Median :0.1000                          Median : 5.000   Median :2012
##  Mean   :0.3063                          Mean   : 5.697   Mean   :2012
##  3rd Qu.:0.5000                          3rd Qu.: 8.000   3rd Qu.:2016
##  Max.   :1.0000                          Max.   :12.000   Max.   :2021
##
##  Chla_uncorrected_ugL     DO_mgL         NO23_mgL     Orthophosphate_mgL
##  Min.   : NA          Min.   : 3.51   Min.   :0.033   Min.   :0.00200
##  1st Qu.: NA          1st Qu.:11.10   1st Qu.:0.676   1st Qu.:0.02310
##  Median : NA          Median :15.80   Median :1.130   Median :0.04730
##  Mean   :NaN          Mean   :32.66   Mean   :1.576   Mean   :0.07033
##  3rd Qu.: NA          3rd Qu.:54.50   3rd Qu.:2.130   3rd Qu.:0.10000
```

```
##  Max.   : NA          Max.   :81.00   Max.   :9.360   Max.   :0.58800
##  NA's   :287         NA's   :114     NA's   :30      NA's   :60
##        pH           TKN_mgL          TP_mgL             NA
##  Min.   :7.100   Min.   :0.503   Min.   :0.0200   Min.   :  0.193
##  1st Qu.:8.100   1st Qu.:1.210   1st Qu.:0.1017   1st Qu.:  5.065
##  Median :8.300   Median :1.465   Median :0.1590   Median : 13.800
##  Mean   :8.272   Mean   :1.542   Mean   :0.1852   Mean   : 17.578
##  3rd Qu.:8.480   3rd Qu.:1.808   3rd Qu.:0.2310   3rd Qu.: 22.300
##  Max.   :9.210   Max.   :4.530   Max.   :0.8420   Max.   :190.000
##  NA's   :76      NA's   :37      NA's   :23       NA's   :180
##     NH3_mgL        Chl_probe_RFU        TSS_mgL          TDP_mgL
##  Min.   :0.01500   Min.   :  0.566   Min.   :  2.00   Min.   :0.013
##  1st Qu.:0.02475   1st Qu.:  4.570   1st Qu.: 10.00   1st Qu.:0.017
##  Median :0.03420   Median : 12.550   Median : 22.30   Median :0.021
##  Mean   :0.06608   Mean   : 25.499   Mean   : 30.11   Mean   :0.021
##  3rd Qu.:0.05850   3rd Qu.: 35.950   3rd Qu.: 37.90   3rd Qu.:0.025
##  Max.   :0.51900   Max.   :173.000   Max.   :385.00   Max.   :0.029
##  NA's   :168       NA's   :199       NA's   :48       NA's   :285
##  MonitoringLocationName MonitoringLocationDescriptionText HUCEightDigitCode
##  Length:287             Length:287                        Length:287
##  Class :character       Class :character                  Class :character
##  Mode  :character       Mode  :character                  Mode  :character
##
##
##
##
##  LatitudeMeasure LongitudeMeasure
##  Min.   :44.11   Min.   :-87.72
##  1st Qu.:44.11   1st Qu.:-87.72
##  Median :44.11   Median :-87.72
##  Mean   :44.11   Mean   :-87.72
##  3rd Qu.:44.11   3rd Qu.:-87.72
##  Max.   :44.11   Max.   :-87.72
##
```

```
ggplot(ManitowocWQ_wide, aes(x = TP_mgL)) +
  geom_histogram(fill = "gray", color = "black")
```

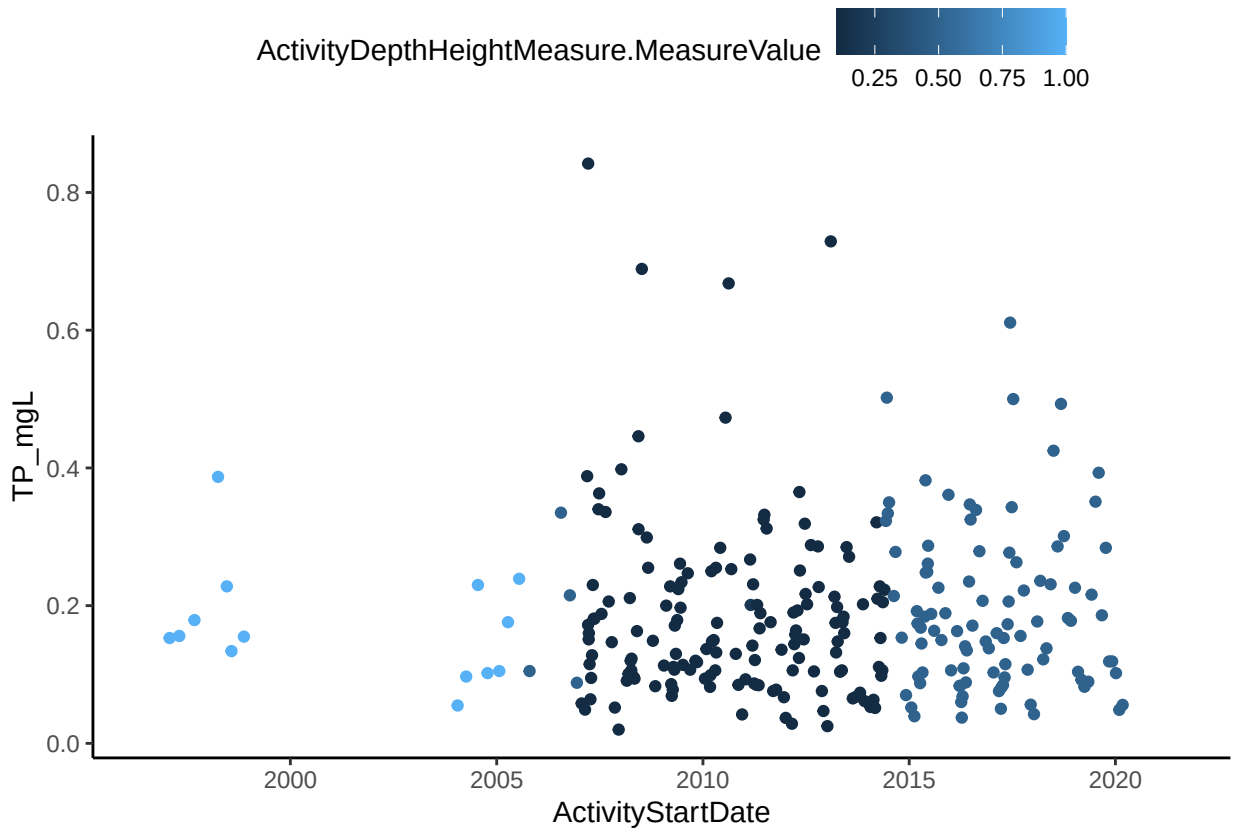
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 23 rows containing non-finite values (stat_bin).
```



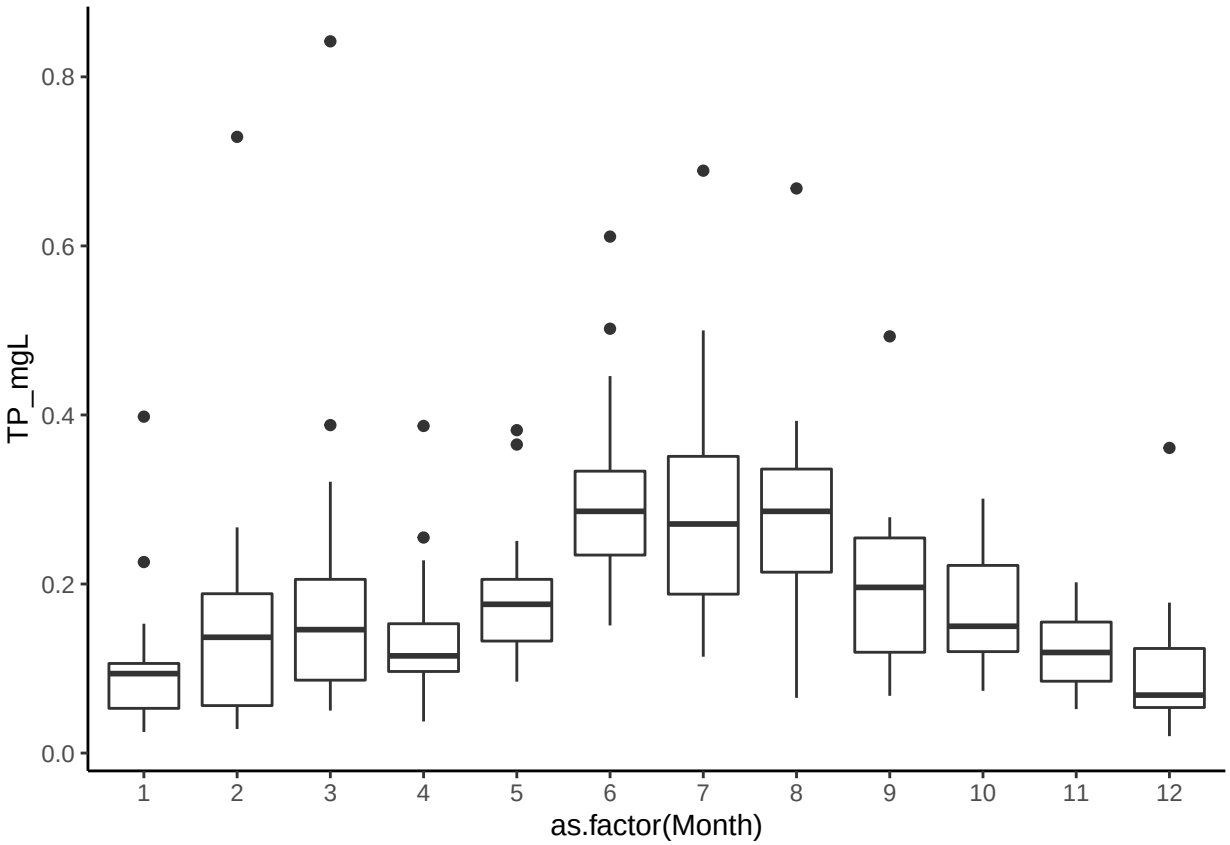

```
ggplot(ManitowocWQ_wide, aes(x = ActivityStartDate, y = TP_mgL, color = ActivityDepthHeightMeasure.Measu
  geom_point() +
  theme(legend.position = "top")
```

```
## Warning: Removed 23 rows containing missing values (geom_point).
```



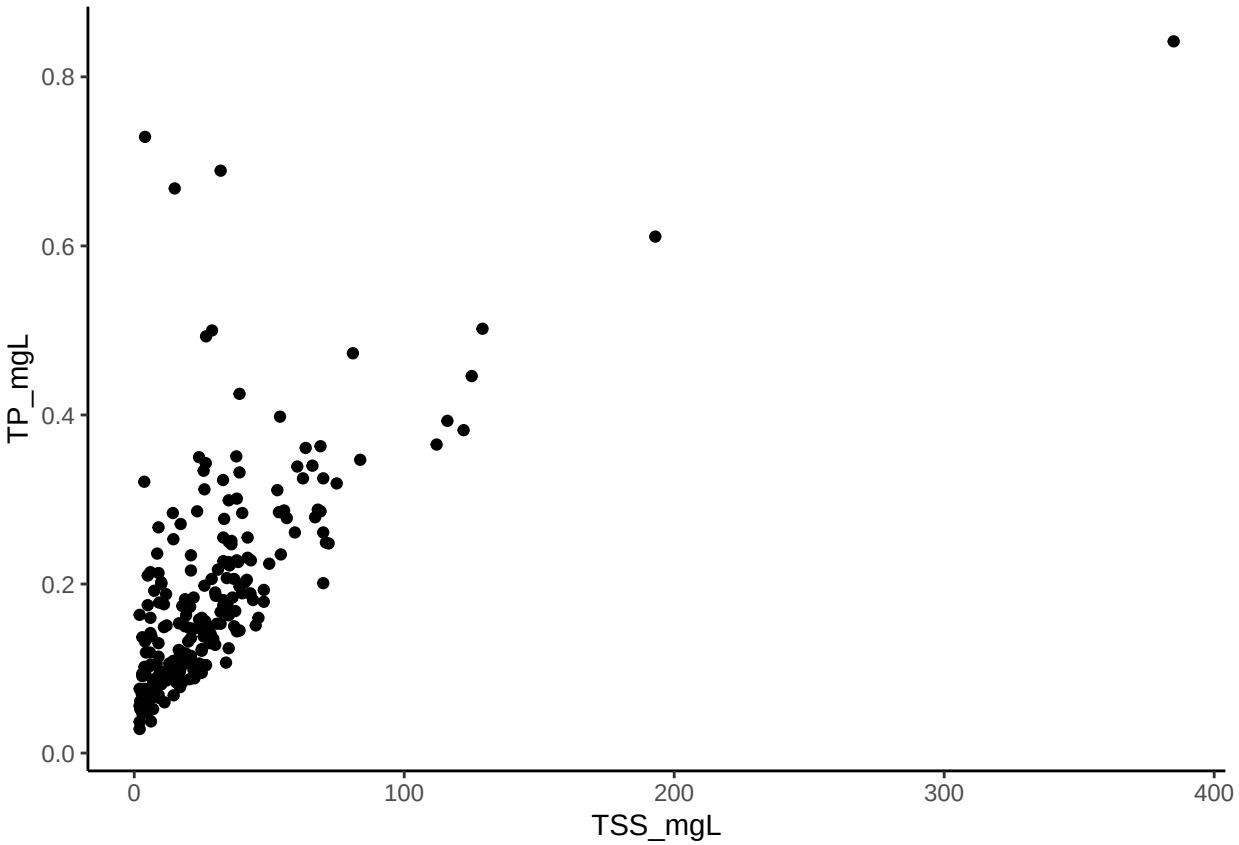

```
ggplot(ManitowocWQ_wide, aes(x = as.factor(Month), y = TP_mgL)) +
  geom_boxplot() +
  theme(legend.position = "top")
```

```
## Warning: Removed 23 rows containing non-finite values (stat_boxplot).
```




```
ggplot(ManitowocWQ_wide, aes(x = TSS_mgL, y = TP_mgL)) +
  geom_point() +
  # scale_x_log10() +
  # scale_y_log10() +
  theme(legend.position = "top")
```

```
## Warning: Removed 63 rows containing missing values (geom_point).
```


TP_mgL
0.0
0.2
0.4
0.6
0.8
0
100
200
300
400
TSS_mgL