XML DATA

WORKING WITH XML DATA

INSTALL AND LOAD PACKAGES

```
pacman::p_load(pacman, tidyverse, xml2)
```

- **xml2** - package for working with xml data

GET XML DATA

- so far - saw how to import hierarchical data in XML and JSON
- now - extended example (same principles for JSON data)

- *data* - from https://data.mo.gov/ - Missouri data portal
- *file* - https://data.mo.gov/api/views/vpge-tj3s.xml
- -> will be looking for **sales-tax-rate by county**

Import XML data:

```
# save URL for the dataset we want
URL <- "https://data.mo.gov/api/views/vpge-tj3s/rows.xml"
```

- result: *environ - values*: URL

- result:

environ - values: URL

Values	
URL	"https://data.mo.gov/api/views/vpge-tj3s/rows.xml"

```
dat <- URL %>%
    read_xml() %>%
    as-list()
```

- result:

envir - data: dat large list (5.2 Mb)
click on it to view structure (hierarchical)

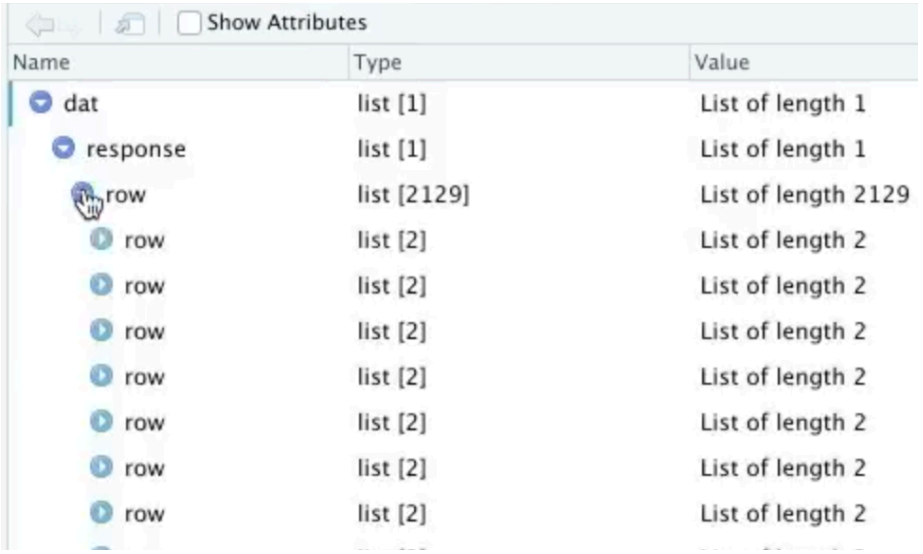
dat	Large list (5.2 Mb)

```
str(dat)   # check structure
```

- result:

console - see that is nested (hierarchical structure)

```
  .. .. ..$ county  :List of 1
  .. .. .. ..$ : chr "Jefferson County             Antonia Fire Protection District And
 Big River Ambulance District"
  .. .. ..$ salestax:List of 1
  .. .. .. ..$ : chr "0.0735"
  .. .. ..- attr(*, "_id")= chr "row-vugk_jz7p.w5ds"
  .. .. ..- attr(*, "_uuid")= chr "00000000-0000-0000-2995-3AA5DF9AB116"
  .. .. ..- attr(*, "_position")= chr "0"
  .. .. ..- attr(*, "_address")= chr "https://data.mo.gov/resource/vpge-tj3s/row-vugk_j
z7p.w5ds"
  .. .. [list output truncated]
```


Show Attributes
Name
Type
Value
dat
list [1]
List of length 1
response
list [1]
List of length 1
row
list [2129]
List of length 2129
row
list [2]
List of length 2
row
list [2]
List of length 2
row
list [2]
List of length 2
row
list [2]
List of length 2
row
list [2]
List of length 2
row
list [2]
List of length 2
row
list [2]
List of length 2


dat
list [1]
List of length 1
response
list [1]
List of length 1
row
list [2129]
List of length 2129
row
list [2]
List of length 2
row
list [2]
List of length 2
county
list [1]
List of length 1
[[1]]
character [1]
'Adair County'
salestax
list [1]
List of length 1
[[1]]
character [1]
'0.056'
row
list [2]
List of length 2
row
list [2]
List of length 2

EXTRACT & COMBINE DATA

start - create tibble > then start unnesting the data

unnest_wider - to get list of counties and taxes:

```
df <- tibble(taxes = dat) %>%
    unnest_wider(taxes) %>%
    print()
```

- result:

environment - *data*: df 1 obs of 1 variable (bc is list)

df	1 obs. of 1 variable

console: list with over 2000 rows of data

```
# A tibble: 1 x 1
  row
  <list>
1 <named list [2,129]>
```

unnest-longer - to get each county/tax pair in a row:

```
df <- df %>%
    unnest_longer(row) %>%
    print()
```

- results:

environment - *data*: df 2019 obs of 2 variables

df	2129 obs. of 2 variables

console: tibble 2,129 x 2
row (has lists) / row-id

```
# A tibble: 2,129 x 2
   row               row_id
   <named list>      <chr>
 1 <named list [2]> row
 2 <named list [2]> row
 3 <named list [2]> row
 4 <named list [2]> row
 5 <named list [2]> row
 6 <named list [2]> row
 7 <named list [2]> row
```

unnest-wider - to get county and tax as separate variables:

```
df <- df %>%
    unnest_wider(row) %>%
    select(-row_id)    # drop unneeded ID
    print()
```

- results:

still 2 lists - have 2 title now (county / salestax)

df 2129 obs. of 2 variables

```
# A tibble: 2,129 x 2
  county     salestax
  <list>     <list>
1 <list [1]> <list [1]>
2 <list [1]> <list [1]>
3 <list [1]> <list [1]>
4 <list [1]> <list [1]>
5 <list [1]> <list [1]>
6 <list [1]> <list [1]>
```

convert from list to character and numeric:

- "**`str_squich`**" removes repeated white space in the county column

```
df <- df %>%
    unnest(county) %>%
    unnest(salestax) %>%
    unnest(salestax) %>%
    mutate(salestax = as.numeric(salestax)) %>%
    mutate(county = str_squich(county)) %>%
    print()
```

- result:

tibble 2,129 x 2
county <chr> / salestax <dbl> = rectangular structure we want!

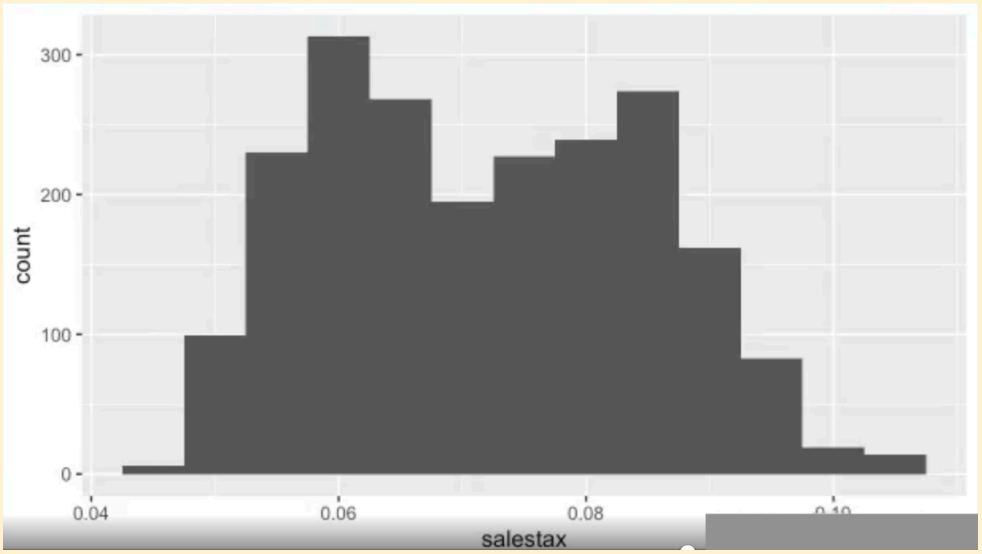
```
# A tibble: 2,129 x 2
   county                                      salestax
   <chr>                                          <dbl>
 1 Adair County                                  0.056
 2 Andrew County                                 0.0592
 3 Andrew County Andrew County Ambulance District 0.0642
 4 Atchison County                               0.0648
 5 Audrain County                                0.0635
 6 Audrain County Audrain Ambulance District     0.0685
 7 Audrain County Van Far Ambulance District     0.0685
 8 Barry County                                  0.0572
 9 Barton County                                 0.0572
10 Bates County                                  0.0522
# … with 2,119 more rows
```

Graph sales tax rates:

```
df %>%
    ggplot(aes(salestax)) +
    geom_histogram(binwidth = 0.005)
```

- result:

a bit bimodal - a lot at 6% and others around 8%



Summary statistics for sales tax rates:

```
df %>% select(salestax) %>% summary()
```

- results:

console:median is 7.25% , mean is close to that

```
    salestax
Min.   :0.04725
1st Qu.:0.06225
Median :0.07225
Mean   :0.07257
3rd Qu.:0.08350
Max.   :0.10679
```