Stat Project Draft Pre-Processing

Christine Nguyen and Salisa Almeida

2024-04-10

Data Loading and Cleaning

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.4
## v ggplot2   3.5.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
theme_set(theme_bw())

df <- read.csv("Real_Estate_Conveyance_Tax_by_Town_20240410.csv")
df |>
  glimpse()
```

```
## Rows: 2,052
## Columns: 24
```

```
## $ Fiscal.Year                                                    <chr> "FY 202~
## $ Town.Code                                                      <int> 1, 2, 3~
## $ Municipality                                                   <chr> "ANDOVE~
## $ Total.Consideration.for.Taxable.Conveyances                    <dbl> 1371529~
## $ Total.Amount.Due                                               <dbl> 104657.~
## $ Consideration.for.Unimproved.Land                              <dbl> 871100,~
## $ Tax.on.Unimproved.Land                                         <dbl> 6533.25~
## $ Total.Consideration.for.Residential.Dwelling                   <dbl> 1248419~
## $ Consideration.for.Residential.Dwelling.Under.Threshold         <dbl> 1248419~
## $ Tax.on.Consideration.Under.Threshold.Residential.Dwelling      <dbl> 93631.4~
## $ Consideration.for.Residential.Dwelling.Over.Threshold          <dbl> NA, NA,~
## $ Tax.on.Consideration.Over.Threshold.Residential.Dwelling       <dbl> NA, NA,~
## $ Consideration.for.Residential.Dwelling.Between..800k..2.5M     <dbl> 0, 1425~
## $ Tax.on.Consideration.Between..800k..2.5M.Residential.Dwelling  <dbl> 0.00, 1~
## $ Consideration.for.Residential.Dwelling.Over..2.5M.Threshold    <dbl> 0, 0, 2~
## $ Tax.on.Consideration.Over..2.5M.Threshold.Residential.Dwelling <dbl> 0.00, 0~
## $ Residential.Property.Other.Than.Dwelling                       <dbl> 0.00, 5~
## $ Tax.on.Residential.Property.Other.Than.Dwelling                <dbl> 0.00, 3~
## $ Nonresidential.Property.Other.Than.Unimproved.Land             <dbl> 360000,~
## $ Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land      <dbl> 4500.00~
## $ Delinquent.Mortgage.Conveyance                                 <dbl> 0.00, 0~
## $ Tax.on.Delinquent.Mortgagor                                    <dbl> 0.00, 0~
## $ Number.of.Taxable.Conveyances                                  <int> 46, 231~
## $ Number.of.Non.Taxable.and.Exempt.Conveyances                   <int> 16, 168~
```

What are all our columns?

```
column_list <- colnames(df)
column_list
```

```
##  [1] "Fiscal.Year"
##  [2] "Town.Code"
##  [3] "Municipality"
##  [4] "Total.Consideration.for.Taxable.Conveyances"
##  [5] "Total.Amount.Due"
##  [6] "Consideration.for.Unimproved.Land"
##  [7] "Tax.on.Unimproved.Land"
##  [8] "Total.Consideration.for.Residential.Dwelling"
##  [9] "Consideration.for.Residential.Dwelling.Under.Threshold"
## [10] "Tax.on.Consideration.Under.Threshold.Residential.Dwelling"
## [11] "Consideration.for.Residential.Dwelling.Over.Threshold"
## [12] "Tax.on.Consideration.Over.Threshold.Residential.Dwelling"
## [13] "Consideration.for.Residential.Dwelling.Between..800k..2.5M"
## [14] "Tax.on.Consideration.Between..800k..2.5M.Residential.Dwelling"
## [15] "Consideration.for.Residential.Dwelling.Over..2.5M.Threshold"
## [16] "Tax.on.Consideration.Over..2.5M.Threshold.Residential.Dwelling"
## [17] "Residential.Property.Other.Than.Dwelling"
## [18] "Tax.on.Residential.Property.Other.Than.Dwelling"
## [19] "Nonresidential.Property.Other.Than.Unimproved.Land"
## [20] "Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land"
## [21] "Delinquent.Mortgage.Conveyance"
## [22] "Tax.on.Delinquent.Mortgagor"
## [23] "Number.of.Taxable.Conveyances"
## [24] "Number.of.Non.Taxable.and.Exempt.Conveyances"
```

Which columns have missing values?

```r
colSums(is.na(df))
```

```
##                                                  Fiscal.Year
##                                                            0
##                                                    Town.Code
##                                                           20
##                                                 Municipality
##                                                            0
##                    Total.Consideration.for.Taxable.Conveyances
##                                                            0
##                                                Total.Amount.Due
##                                                            0
##                             Consideration.for.Unimproved.Land
##                                                            0
##                                           Tax.on.Unimproved.Land
##                                                            0
##                   Total.Consideration.for.Residential.Dwelling
##                                                            0
##         Consideration.for.Residential.Dwelling.Under.Threshold
##                                                            0
##      Tax.on.Consideration.Under.Threshold.Residential.Dwelling
##                                                            0
##          Consideration.for.Residential.Dwelling.Over.Threshold
##                                                          513
##       Tax.on.Consideration.Over.Threshold.Residential.Dwelling
##                                                          513
##     Consideration.for.Residential.Dwelling.Between..800k..2.5M
##                                                         1539
##  Tax.on.Consideration.Between..800k..2.5M.Residential.Dwelling
##                                                         1539
##    Consideration.for.Residential.Dwelling.Over..2.5M.Threshold
##                                                         1539
## Tax.on.Consideration.Over..2.5M.Threshold.Residential.Dwelling
##                                                         1539
##                         Residential.Property.Other.Than.Dwelling
##                                                            0
##                 Tax.on.Residential.Property.Other.Than.Dwelling
##                                                            0
##              Nonresidential.Property.Other.Than.Unimproved.Land
##                                                            0
##      Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land
##                                                            0
##                                    Delinquent.Mortgage.Conveyance
##                                                            0
##                                      Tax.on.Delinquent.Mortgagor
##                                                            0
##                                     Number.of.Taxable.Conveyances
##                                                            0
##                    Number.of.Non.Taxable.and.Exempt.Conveyances
##                                                            0
```

Removing columns missing 75% of data.

```r
# 75% of the data is missing in these columns so we will not impute them
# imputing would create a big bias so we remove the columns
```

```r
df <- df |>
  select(-Consideration.for.Residential.Dwelling.Between..800k..2.5M,
         -Tax.on.Consideration.Between..800k..2.5M.Residential.Dwelling,
         -Consideration.for.Residential.Dwelling.Over..2.5M.Threshold,
         -Tax.on.Consideration.Over..2.5M.Threshold.Residential.Dwelling)

head(df)
```

```
##   Fiscal.Year Town.Code       Municipality
## 1  FY 2022-23         1      ANDOVER (001)
## 2  FY 2022-23         2      ANSONIA (002)
## 3  FY 2022-23         3      ASHFORD (003)
## 4  FY 2022-23         4         AVON (004)
## 5  FY 2022-23         5  BARKHAMSTED (005)
## 6  FY 2022-23         6 BEACON FALLS (006)
##   Total.Consideration.for.Taxable.Conveyances Total.Amount.Due
## 1                                    13715290         104657.2
## 2                                    72506335         559369.4
## 3                                    27031172         222373.3
## 4                                   232014178        1809584.4
## 5                                    23752400         188493.0
## 6                                    45074056         340780.4
##   Consideration.for.Unimproved.Land Tax.on.Unimproved.Land
## 1                            871100                6533.25
## 2                           2376510               12536.25
## 3                            981140                7358.55
## 4                           6335375               47515.31
## 5                           1291900                9539.25
## 6                            616000                4620.00
##   Total.Consideration.for.Residential.Dwelling
## 1                                     12484190
## 2                                     64892327
## 3                                     24446812
## 4                                    185546999
## 5                                     20200500
## 6                                     43913056
##   Consideration.for.Residential.Dwelling.Under.Threshold
## 1                                               12484190
## 2                                               63467327
## 3                                               21994812
## 4                                              169940173
## 5                                               20100500
## 6                                               43913056
##   Tax.on.Consideration.Under.Threshold.Residential.Dwelling
## 1                                                   93631.43
## 2                                                  473852.47
## 3                                                  161646.09
## 4                                                 1243943.74
## 5                                                  150753.74
## 6                                                  329347.95
##   Consideration.for.Residential.Dwelling.Over.Threshold
## 1                                                    NA
## 2                                                    NA
## 3                                                    NA
```

```
## 4                                                    NA
## 5                                                    NA
## 6                                                    NA
##   Tax.on.Consideration.Over.Threshold.Residential.Dwelling
## 1                                                       NA
## 2                                                       NA
## 3                                                       NA
## 4                                                       NA
## 5                                                       NA
## 6                                                       NA
##   Residential.Property.Other.Than.Dwelling
## 1                                     0.00
## 2                                 50000.00
## 3                                  8320.23
## 4                                34856500.00
## 5                                  260000.00
## 6                                     0.00
##   Tax.on.Residential.Property.Other.Than.Dwelling
## 1                                             0.0
## 2                                           375.0
## 3                                            62.4
## 4                                          261423.8
## 5                                           1950.0
## 6                                             0.0
##   Nonresidential.Property.Other.Than.Unimproved.Land
## 1                                             360000
## 2                                            5187497
## 3                                            1594900
## 4                                            5275304
## 5                                            2000000
## 6                                             545000
##   Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land
## 1                                                   4500.00
## 2                                                  54793.72
## 3                                                  19936.25
## 4                                                  65941.30
## 5                                                  25000.00
## 6                                                   6812.50
##   Delinquent.Mortgage.Conveyance Tax.on.Delinquent.Mortgagor
## 1                              0                           0
## 2                              0                           0
## 3                              0                           0
## 4                              0                           0
## 5                              0                           0
## 6                              0                           0
##   Number.of.Taxable.Conveyances Number.of.Non.Taxable.and.Exempt.Conveyances
## 1                            46                                          16
## 2                           231                                         168
## 3                            84                                          56
## 4                           392                                         205
## 5                            76                                          35
## 6                           132                                          80
```

now we look at missing rows for Town.Code

```r
# double checking town code missing values by rows
rows_with_missing_values <- which(!complete.cases(df$Town.Code))

df[rows_with_missing_values, ]
```

```
##      Fiscal.Year Town.Code       Municipality
## 171   FY 2022-23        NA ALL MUNICIPALITIES
## 342   FY 2021-22        NA ALL MUNICIPALITIES
## 513   FY 2020-21        NA ALL MUNICIPALITIES
## 684   FY 2019-20        NA ALL MUNICIPALITIES
## 854   FY 2018-19        NA       TOWN UNKNOWN
## 855   FY 2018-19        NA ALL MUNICIPALITIES
## 1025  FY 2017-18        NA       TOWN UNKNOWN
## 1026  FY 2017-18        NA ALL MUNICIPALITIES
## 1196  FY 2016-17        NA       TOWN UNKNOWN
## 1197  FY 2016-17        NA ALL MUNICIPALITIES
## 1367  FY 2015-16        NA       TOWN UNKNOWN
## 1368  FY 2015-16        NA ALL MUNICIPALITIES
## 1538  FY 2014-15        NA       TOWN UNKNOWN
## 1539  FY 2014-15        NA ALL MUNICIPALITIES
## 1709  FY 2013-14        NA       TOWN UNKNOWN
## 1710  FY 2013-14        NA ALL MUNICIPALITIES
## 1880  FY 2012-13        NA       TOWN UNKNOWN
## 1881  FY 2012-13        NA ALL MUNICIPALITIES
## 2051  FY 2011-12        NA       TOWN UNKNOWN
## 2052  FY 2011-12        NA ALL MUNICIPALITIES
##      Total.Consideration.for.Taxable.Conveyances Total.Amount.Due
## 171                                  30747530560      281845782.6
## 342                                  40430473828      368892914.9
## 513                                  36450471819      332634075.0
## 684                                  22822275707      199101702.0
## 854                                     32510743         274646.0
## 855                                  21734674444      188582092.0
## 1025                                    41428954         338477.0
## 1026                                 21053727611      183480574.0
## 1196                                    39895583         311402.0
## 1197                                 20769287150      181932050.0
## 1367                                    13718784         113666.0
## 1368                                 19514729875      172919212.0
## 1538                                    12806879         106689.0
## 1539                                 18099190110      161469219.0
## 1709                                    31898671         295427.1
## 1710                                 17656454171      157237245.6
## 1880                                    34636480         310495.0
## 1881                                 15701355899      138700078.0
## 2051                                    27974509         243497.0
## 2052                                 13624513832      119354168.0
##      Consideration.for.Unimproved.Land Tax.on.Unimproved.Land
## 171                          777945111               5647639
## 342                         1119890577               7302463
## 513                          832440973               6243307
## 684                          572083446               4290626
## 854                            1952000                 14640
## 855                          612280928               4592107
```

```
## 1025                          683500                 5126
## 1026                       579744060                4348080
## 1196                         2335863                  17519
## 1197                       504013379                3780100
## 1367                          487000                   3653
## 1368                       470055687                3525418
## 1538                          281000                   2108
## 1539                       467342711                3505070
## 1709                          374000                   2805
## 1710                       509135539                3818517
## 1880                          645800                   4844
## 1881                       500076406                3750573
## 2051                          266400                   1998
## 2052                       505708322                3792812
##      Total.Consideration.for.Residential.Dwelling
## 171                                  24199965543
## 342                                  30661825749
## 513                                  31509764385
## 684                                  18091852342
## 854                                     28095743
## 855                                  17092333988
## 1025                                    38745453
## 1026                                 16815621696
## 1196                                    34751120
## 1197                                 16382038360
## 1367                                    11931784
## 1368                                 15566885112
## 1538                                    12010879
## 1539                                 14333849858
## 1709                                    27521262
## 1710                                 14200309062
## 1880                                    32816180
## 1881                                 12963640399
## 2051                                    27244412
## 2052                                 11107424066
##      Consideration.for.Residential.Dwelling.Under.Threshold
## 171                                       19507012974
## 342                                       24698532611
## 513                                       25184570625
## 684                                       15577764890
## 854                                          24345743
## 855                                       14725654952
## 1025                                         35193453
## 1026                                      14217212746
## 1196                                         32824120
## 1197                                      13830676451
## 1367                                         11076784
## 1368                                      12981867329
## 1538                                         10023379
## 1539                                      11494975125
## 1709                                         20287262
## 1710                                      11361660948
## 1880                                         23840462
## 1881                                      10504796201
```

```
## 2051                                                  20970412
## 2052                                                 9147821844
##      Tax.on.Consideration.Under.Threshold.Residential.Dwelling
## 171                                                  144448308.2
## 342                                                  183635829.6
## 513                                                  188884280.0
## 684                                                  116833237.0
## 854                                                     182593.0
## 855                                                  110441065.0
## 1025                                                    263951.0
## 1026                                                 106629096.0
## 1196                                                    246181.0
## 1197                                                 103730073.0
## 1367                                                     83076.0
## 1368                                                  97364005.0
## 1538                                                     75175.0
## 1539                                                  86212313.0
## 1709                                                    152154.5
## 1710                                                  85212457.1
## 1880                                                    178803.0
## 1881                                                  78785972.0
## 2051                                                    157278.0
## 2052                                                  68600148.0
##      Consideration.for.Residential.Dwelling.Over.Threshold
## 171                                                     NA
## 342                                                     NA
## 513                                                     NA
## 684                                                2514087452
## 854                                                   3750000
## 855                                                2366679036
## 1025                                                  3552000
## 1026                                               2598408950
## 1196                                                  1927000
## 1197                                               2551361908
## 1367                                                   855000
## 1368                                               2585017783
## 1538                                                  1987500
## 1539                                               2838874733
## 1709                                                  7234000
## 1710                                               2838648114
## 1880                                                  8975718
## 1881                                               2458844199
## 2051                                                  6274000
## 2052                                               1959602222
##      Tax.on.Consideration.Over.Threshold.Residential.Dwelling
## 171                                                        NA
## 342                                                        NA
## 513                                                        NA
## 684                                                  31426093
## 854                                                     46875
## 855                                                  29583488
## 1025                                                    44400
## 1026                                                 32480112
## 1196                                                    24088
```

```
## 1197                                                    31892024
## 1367                                                       10688
## 1368                                                    32312722
## 1538                                                       24844
## 1539                                                    35485934
## 1709                                                       90425
## 1710                                                    35483101
## 1880                                                      112196
## 1881                                                    30735552
## 2051                                                       78425
## 2052                                                    24494028
##      Residential.Property.Other.Than.Dwelling
## 171                                1766408961
## 342                                3146134321
## 513                                1391739739
## 684                                1083383313
## 854                                     50000
## 855                                1278831709
## 1025                                        0
## 1026                               1113208506
## 1196                                  2298600
## 1197                               1142552221
## 1367                                        0
## 1368                                727955159
## 1538                                   375000
## 1539                                983785468
## 1709                                        0
## 1710                                814600505
## 1880                                     6000
## 1881                                498035852
## 2051                                        0
## 2052                                493254546
##      Tax.on.Residential.Property.Other.Than.Dwelling
## 171                                         12396441
## 342                                         20913725
## 513                                         10438048
## 684                                          8125375
## 854                                              375
## 855                                          9591238
## 1025                                               0
## 1026                                         8349064
## 1196                                           17240
## 1197                                         8569142
## 1367                                               0
## 1368                                         5459664
## 1538                                            2813
## 1539                                         7378391
## 1709                                               0
## 1710                                         6109504
## 1880                                              45
## 1881                                         3735269
## 2051                                               0
## 2052                                         3699409
##      Nonresidential.Property.Other.Than.Unimproved.Land
```

```
## 171                                                3992789259
## 342                                                5500621293
## 513                                                2716526721
## 684                                                3072839354
## 854                                                   2413000
## 855                                                2747997161
## 1025                                                  2000001
## 1026                                               2517114447
## 1196                                                   510000
## 1197                                               2681117417
## 1367                                                  1300000
## 1368                                               2726729801
## 1538                                                   140000
## 1539                                               2306183840
## 1709                                                  4003409
## 1710                                               2124119748
## 1880                                                  1168500
## 1881                                               1729137578
## 2051                                                   463698
## 2052                                               1476435544
##      Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land
## 171                                                48561887.74
## 342                                                66968322.81
## 513                                                33956584.00
## 684                                                38410492.00
## 854                                                   30163.00
## 855                                                34349965.00
## 1025                                                  25000.00
## 1026                                               31463931.00
## 1196                                                   6375.00
## 1197                                               33513968.00
## 1367                                                  16250.00
## 1368                                               34084123.00
## 1538                                                   1750.00
## 1539                                               28827298.00
## 1709                                                  50042.61
## 1710                                               26551496.84
## 1880                                                  14606.00
## 1881                                               21614220.00
## 2051                                                   5796.00
## 2052                                               18455444.00
##      Delinquent.Mortgage.Conveyance Tax.on.Delinquent.Mortgagor
## 171                        10421687                    59969.73
## 342                         2001888                    13453.79
## 513                               1                        0.00
## 684                         2117252                    15879.00
## 854                               0                        0.00
## 855                         3230658                    24230.00
## 1025                              0                        0.00
## 1026                       28038901                   210292.00
## 1196                              0                        0.00
## 1197                       59565774                   446743.00
## 1367                              0                        0.00
## 1368                       23104116                   173281.00
```

```
## 1538                            0                      0.00
## 1539                      8028233                  60212.00
## 1709                            0                      0.00
## 1710                      8289318                  62169.88
## 1880                            0                      0.00
## 1881                     10465663                  78492.00
## 2051                            0                      0.00
## 2052                     41691355                 312326.00
##      Number.of.Taxable.Conveyances Number.of.Non.Taxable.and.Exempt.Conveyances
## 171                          56135                                        34231
## 342                          75504                                        36948
## 513                          79065                                        30279
## 684                          57556                                        19432
## 854                            106                                           58
## 855                          56654                                        27686
## 1025                           136                                           83
## 1026                         53983                                        33809
## 1196                           139                                           91
## 1197                         53387                                        35111
## 1367                           164                                           40
## 1368                         49975                                        37196
## 1538                            42                                           32
## 1539                         43235                                        36235
## 1709                            68                                           52
## 1710                         42291                                        36027
## 1880                            94                                            1
## 1881                         40177                                        35772
## 2051                            83                                           70
## 2052                         36384                                        35951
```

```
# town code does not actually have missing values, it is just coded differently.
# data dictionary says "Town Unknown" is coded as 000
# "All Municipalities" had no code so I created 170 as the largest code is 169
# fill in Town.Code based on conditions
df <- df %>%
  mutate(Town.Code = case_when(
    Municipality == "ALL MUNICIPALITIES" ~ "170",
    Municipality == "TOWN UNKNOWN" ~ "0",
    TRUE ~ as.character(Town.Code)  # Keep the original value for other cases
  ))

head(df)
```

```
##   Fiscal.Year Town.Code       Municipality
## 1  FY 2022-23         1      ANDOVER (001)
## 2  FY 2022-23         2      ANSONIA (002)
## 3  FY 2022-23         3      ASHFORD (003)
## 4  FY 2022-23         4         AVON (004)
## 5  FY 2022-23         5  BARKHAMSTED (005)
## 6  FY 2022-23         6 BEACON FALLS (006)
##   Total.Consideration.for.Taxable.Conveyances Total.Amount.Due
## 1                                    13715290         104657.2
## 2                                    72506335         559369.4
## 3                                    27031172         222373.3
## 4                                   232014178        1809584.4
```

```
## 5                                    23752400         188493.0
## 6                                    45074056         340780.4
##   Consideration.for.Unimproved.Land Tax.on.Unimproved.Land
## 1                            871100                6533.25
## 2                           2376510               12536.25
## 3                            981140                7358.55
## 4                           6335375               47515.31
## 5                           1291900                9539.25
## 6                            616000                4620.00
##   Total.Consideration.for.Residential.Dwelling
## 1                                     12484190
## 2                                     64892327
## 3                                     24446812
## 4                                    185546999
## 5                                     20200500
## 6                                     43913056
##   Consideration.for.Residential.Dwelling.Under.Threshold
## 1                                               12484190
## 2                                               63467327
## 3                                               21994812
## 4                                              169940173
## 5                                               20100500
## 6                                               43913056
##   Tax.on.Consideration.Under.Threshold.Residential.Dwelling
## 1                                                  93631.43
## 2                                                 473852.47
## 3                                                 161646.09
## 4                                                1243943.74
## 5                                                 150753.74
## 6                                                 329347.95
##   Consideration.for.Residential.Dwelling.Over.Threshold
## 1                                                    NA
## 2                                                    NA
## 3                                                    NA
## 4                                                    NA
## 5                                                    NA
## 6                                                    NA
##   Tax.on.Consideration.Over.Threshold.Residential.Dwelling
## 1                                                       NA
## 2                                                       NA
## 3                                                       NA
## 4                                                       NA
## 5                                                       NA
## 6                                                       NA
##   Residential.Property.Other.Than.Dwelling
## 1                                     0.00
## 2                                 50000.00
## 3                                  8320.23
## 4                              34856500.00
## 5                                260000.00
## 6                                     0.00
##   Tax.on.Residential.Property.Other.Than.Dwelling
## 1                                             0.0
## 2                                           375.0
```

```
## 3                                          62.4
## 4                                        261423.8
## 5                                          1950.0
## 6                                             0.0
##   Nonresidential.Property.Other.Than.Unimproved.Land
## 1                                             360000
## 2                                            5187497
## 3                                            1594900
## 4                                            5275304
## 5                                            2000000
## 6                                             545000
##   Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land
## 1                                                   4500.00
## 2                                                  54793.72
## 3                                                  19936.25
## 4                                                  65941.30
## 5                                                  25000.00
## 6                                                   6812.50
##   Delinquent.Mortgage.Conveyance Tax.on.Delinquent.Mortgagor
## 1                              0                           0
## 2                              0                           0
## 3                              0                           0
## 4                              0                           0
## 5                              0                           0
## 6                              0                           0
##   Number.of.Taxable.Conveyances Number.of.Non.Taxable.and.Exempt.Conveyances
## 1                            46                                          16
## 2                           231                                         168
## 3                            84                                          56
## 4                           392                                         205
## 5                            76                                          35
## 6                           132                                          80
```

```
colSums(is.na(df))
```

```
##                                          Fiscal.Year
##                                                    0
##                                            Town.Code
##                                                    0
##                                         Municipality
##                                                    0
##               Total.Consideration.for.Taxable.Conveyances
##                                                    0
##                                       Total.Amount.Due
##                                                    0
##                         Consideration.for.Unimproved.Land
##                                                    0
##                                    Tax.on.Unimproved.Land
##                                                    0
##              Total.Consideration.for.Residential.Dwelling
##                                                    0
##    Consideration.for.Residential.Dwelling.Under.Threshold
##                                                    0
## Tax.on.Consideration.Under.Threshold.Residential.Dwelling
##                                                    0
```

```
##     Consideration.for.Residential.Dwelling.Over.Threshold
##                                                       513
##  Tax.on.Consideration.Over.Threshold.Residential.Dwelling
##                                                       513
##                   Residential.Property.Other.Than.Dwelling
##                                                         0
##           Tax.on.Residential.Property.Other.Than.Dwelling
##                                                         0
##        Nonresidential.Property.Other.Than.Unimproved.Land
##                                                         0
## Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land
##                                                         0
##                              Delinquent.Mortgage.Conveyance
##                                                         0
##                                 Tax.on.Delinquent.Mortgagor
##                                                         0
##                               Number.of.Taxable.Conveyances
##                                                         0
##              Number.of.Non.Taxable.and.Exempt.Conveyances
##                                                         0
```

creating column that shows which rows have been imputed by mean with the columns missing 25% of values

```
df$Imputed <- ifelse(is.na(df$Consideration.for.Residential.Dwelling.Over.Threshold) | is.na(df$Tax.on.C
```

What are the means of the two columns with 25% missing values?

```
# these columns only have 513 missing values so we will impute with mean, 25% missing
mean(df$Consideration.for.Residential.Dwelling.Over.Threshold, na.rm = TRUE)
```

```
## [1] 29514652
```

```
mean(df$Tax.on.Consideration.Over.Threshold.Residential.Dwelling, na.rm = TRUE)
```

```
## [1] 368931.9
```

implement the mean imputation for these two columns: Imputing with mean for 2 columns missing only 25% of data.

```
df <- df |>
  mutate(Consideration.for.Residential.Dwelling.Over.Threshold =
           ifelse(is.na(Consideration.for.Residential.Dwelling.Over.Threshold),
                  mean(Consideration.for.Residential.Dwelling.Over.Threshold, na.rm = TRUE),
                  Consideration.for.Residential.Dwelling.Over.Threshold))

df <- df |>
  mutate(Tax.on.Consideration.Over.Threshold.Residential.Dwelling =
           ifelse(is.na(Tax.on.Consideration.Over.Threshold.Residential.Dwelling),
                  mean(Tax.on.Consideration.Over.Threshold.Residential.Dwelling, na.rm = TRUE),
                  Tax.on.Consideration.Over.Threshold.Residential.Dwelling))

colSums(is.na(df))
```

```
##                                         Fiscal.Year
##                                                   0
##                                           Town.Code
##                                                   0
##                                        Municipality
```

```
##                                                          0
##               Total.Consideration.for.Taxable.Conveyances
##                                                          0
##                                           Total.Amount.Due
##                                                          0
##                         Consideration.for.Unimproved.Land
##                                                          0
##                                     Tax.on.Unimproved.Land
##                                                          0
##              Total.Consideration.for.Residential.Dwelling
##                                                          0
##    Consideration.for.Residential.Dwelling.Under.Threshold
##                                                          0
## Tax.on.Consideration.Under.Threshold.Residential.Dwelling
##                                                          0
##     Consideration.for.Residential.Dwelling.Over.Threshold
##                                                          0
##  Tax.on.Consideration.Over.Threshold.Residential.Dwelling
##                                                          0
##                  Residential.Property.Other.Than.Dwelling
##                                                          0
##           Tax.on.Residential.Property.Other.Than.Dwelling
##                                                          0
##        Nonresidential.Property.Other.Than.Unimproved.Land
##                                                          0
## Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land
##                                                          0
##                              Delinquent.Mortgage.Conveyance
##                                                          0
##                                 Tax.on.Delinquent.Mortgagor
##                                                          0
##                               Number.of.Taxable.Conveyances
##                                                          0
##              Number.of.Non.Taxable.and.Exempt.Conveyances
##                                                          0
##                                                    Imputed
##                                                          0
```

Data Exploration

```r
par(mfrow = c(9, 2))

# first 10 municipals
group_municipalities <- df %>%
  count(Municipality) %>%
  arrange(desc(n)) %>%
  .[1:10, "Municipality"]

plot1 <- df %>%
  filter(Municipality %in% group_municipalities) %>%
  mutate(log_total_consideration = log(Total.Consideration.for.Taxable.Conveyances),
         Fiscal.Year_numeric = as.numeric(substring(Fiscal.Year, 4, 7))) %>%
  ggplot(aes(x = Fiscal.Year_numeric, y = log_total_consideration, color = Municipality)) +
```

```
  geom_line() +
  geom_point() +
  labs(x = "Fiscal Year",
       y = "Natural Logarithm of Total Consideration for Taxable Conveyances",
       color = "Municipality")

# next 10 municipals
group_municipalities <- df %>%
  count(Municipality) %>%
  arrange(desc(n)) %>%
  .[11:20, "Municipality"]

# Plot with legend trimming
plot2 <- df %>%
  filter(Municipality %in% group_municipalities) %>%
  mutate(log_total_consideration = log(Total.Consideration.for.Taxable.Conveyances),
         Fiscal.Year_numeric = as.numeric(substring(Fiscal.Year, 4, 7))) %>%
  ggplot(aes(x = Fiscal.Year_numeric, y = log_total_consideration, color = Municipality)) +
  geom_line() +
  geom_point() +
  labs(x = "Fiscal Year",
       y = "Natural Logarithm of Total Consideration for Taxable Conveyances",
       color = "Municipality")


# next 10 municipals
group_municipalities <- df %>%
  count(Municipality) %>%
  arrange(desc(n)) %>%
  .[21:30, "Municipality"]

# Plot with legend trimming
plot3 <- df %>%
  filter(Municipality %in% group_municipalities) %>%
  mutate(log_total_consideration = log(Total.Consideration.for.Taxable.Conveyances),
         Fiscal.Year_numeric = as.numeric(substring(Fiscal.Year, 4, 7))) %>%
  ggplot(aes(x = Fiscal.Year_numeric, y = log_total_consideration, color = Municipality)) +
  geom_line() +
  geom_point() +
  labs(x = "Fiscal Year",
       y = "Natural Logarithm of Total Consideration for Taxable Conveyances",
       color = "Municipality")

# next 10 municipals
group_municipalities <- df %>%
  count(Municipality) %>%
  arrange(desc(n)) %>%
  .[31:40, "Municipality"]

# Plot with legend trimming
plot4 <- df %>%
  filter(Municipality %in% group_municipalities) %>%
  mutate(log_total_consideration = log(Total.Consideration.for.Taxable.Conveyances),
```

```
        Fiscal.Year_numeric = as.numeric(substring(Fiscal.Year, 4, 7))) %>%
  ggplot(aes(x = Fiscal.Year_numeric, y = log_total_consideration, color = Municipality)) +
  geom_line() +
  geom_point() +
  labs(x = "Fiscal Year",
       y = "Natural Logarithm of Total Consideration for Taxable Conveyances",
       color = "Municipality")

# next 10 municipals
group_municipalities <- df %>%
  count(Municipality) %>%
  arrange(desc(n)) %>%
  .[41:50, "Municipality"]

# Plot with legend trimming
plot5 <- df %>%
  filter(Municipality %in% group_municipalities) %>%
  mutate(log_total_consideration = log(Total.Consideration.for.Taxable.Conveyances),
         Fiscal.Year_numeric = as.numeric(substring(Fiscal.Year, 4, 7))) %>%
  ggplot(aes(x = Fiscal.Year_numeric, y = log_total_consideration, color = Municipality)) +
  geom_line() +
  geom_point() +
  labs(x = "Fiscal Year",
       y = "Natural Logarithm of Total Consideration for Taxable Conveyances",
       color = "Municipality")

# next 10 municipals
group_municipalities <- df %>%
  count(Municipality) %>%
  arrange(desc(n)) %>%
  .[51:60, "Municipality"]

# Plot with legend trimming
plot6 <- df %>%
  filter(Municipality %in% group_municipalities) %>%
  mutate(log_total_consideration = log(Total.Consideration.for.Taxable.Conveyances),
         Fiscal.Year_numeric = as.numeric(substring(Fiscal.Year, 4, 7))) %>%
  ggplot(aes(x = Fiscal.Year_numeric, y = log_total_consideration, color = Municipality)) +
  geom_line() +
  geom_point() +
  labs(x = "Fiscal Year",
       y = "Natural Logarithm of Total Consideration for Taxable Conveyances",
       color = "Municipality")

# next 10 municipals
group_municipalities <- df %>%
  count(Municipality) %>%
  arrange(desc(n)) %>%
  .[61:70, "Municipality"]

# Plot with legend trimming
plot7 <- df %>%
  filter(Municipality %in% group_municipalities) %>%
```

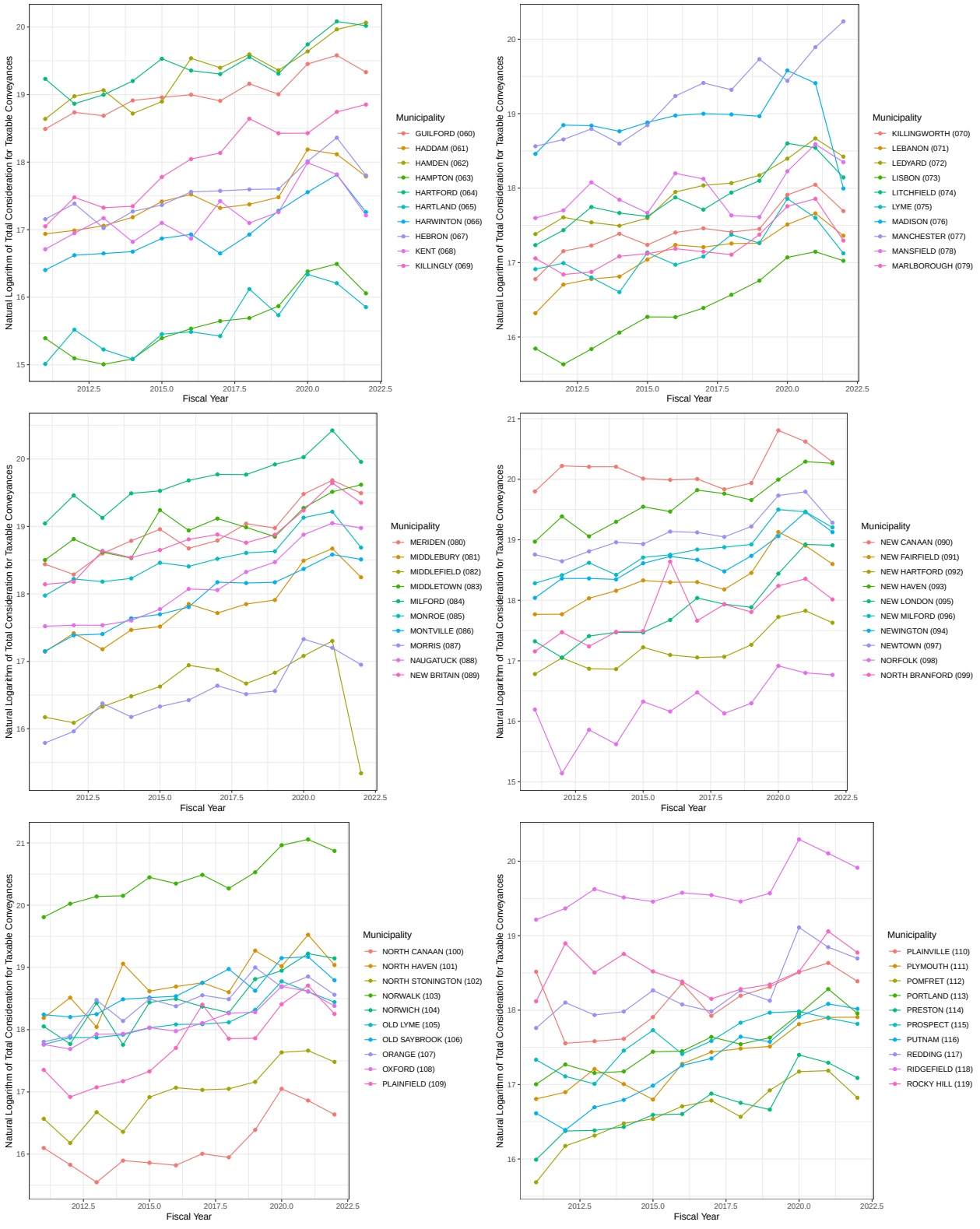
```
  mutate(log_total_consideration = log(Total.Consideration.for.Taxable.Conveyances),
         Fiscal.Year_numeric = as.numeric(substring(Fiscal.Year, 4, 7))) %>%
  ggplot(aes(x = Fiscal.Year_numeric, y = log_total_consideration, color = Municipality)) +
  geom_line() +
  geom_point() +
  labs(x = "Fiscal Year",
       y = "Natural Logarithm of Total Consideration for Taxable Conveyances",
       color = "Municipality")

# next 10 municipals
group_municipalities <- df %>%
  count(Municipality) %>%
  arrange(desc(n)) %>%
  .[71:80, "Municipality"]

# Plot with legend trimming
plot8 <- df %>%
  filter(Municipality %in% group_municipalities) %>%
  mutate(log_total_consideration = log(Total.Consideration.for.Taxable.Conveyances),
         Fiscal.Year_numeric = as.numeric(substring(Fiscal.Year, 4, 7))) %>%
  ggplot(aes(x = Fiscal.Year_numeric, y = log_total_consideration, color = Municipality)) +
  geom_line() +
  geom_point() +
  labs(x = "Fiscal Year",
       y = "Natural Logarithm of Total Consideration for Taxable Conveyances",
       color = "Municipality")

# next 10 municipals
group_municipalities <- df %>%
  count(Municipality) %>%
  arrange(desc(n)) %>%
  .[81:90, "Municipality"]

# Plot with legend trimming
plot9 <- df %>%
  filter(Municipality %in% group_municipalities) %>%
  mutate(log_total_consideration = log(Total.Consideration.for.Taxable.Conveyances),
         Fiscal.Year_numeric = as.numeric(substring(Fiscal.Year, 4, 7))) %>%
  ggplot(aes(x = Fiscal.Year_numeric, y = log_total_consideration, color = Municipality)) +
  geom_line() +
  geom_point() +
  labs(x = "Fiscal Year",
       y = "Natural Logarithm of Total Consideration for Taxable Conveyances",
       color = "Municipality")

# next 10 municipals
group_municipalities <- df %>%
  count(Municipality) %>%
  arrange(desc(n)) %>%
  .[91:100, "Municipality"]

# Plot with legend trimming
plot10 <- df %>%
```

```r
  filter(Municipality %in% group_municipalities) %>%
  mutate(log_total_consideration = log(Total.Consideration.for.Taxable.Conveyances),
         Fiscal.Year_numeric = as.numeric(substring(Fiscal.Year, 4, 7))) %>%
  ggplot(aes(x = Fiscal.Year_numeric, y = log_total_consideration, color = Municipality)) +
  geom_line() +
  geom_point() +
  labs(x = "Fiscal Year",
       y = "Natural Logarithm of Total Consideration for Taxable Conveyances",
       color = "Municipality")

# next 10 municipals
group_municipalities <- df %>%
  count(Municipality) %>%
  arrange(desc(n)) %>%
  .[101:110, "Municipality"]

# Plot with legend trimming
plot11 <- df %>%
  filter(Municipality %in% group_municipalities) %>%
  mutate(log_total_consideration = log(Total.Consideration.for.Taxable.Conveyances),
         Fiscal.Year_numeric = as.numeric(substring(Fiscal.Year, 4, 7))) %>%
  ggplot(aes(x = Fiscal.Year_numeric, y = log_total_consideration, color = Municipality)) +
  geom_line() +
  geom_point() +
  labs(x = "Fiscal Year",
       y = "Natural Logarithm of Total Consideration for Taxable Conveyances",
       color = "Municipality")

# next 10 municipals
group_municipalities <- df %>%
  count(Municipality) %>%
  arrange(desc(n)) %>%
  .[111:120, "Municipality"]

# Plot with legend trimming
plot12 <- df %>%
  filter(Municipality %in% group_municipalities) %>%
  mutate(log_total_consideration = log(Total.Consideration.for.Taxable.Conveyances),
         Fiscal.Year_numeric = as.numeric(substring(Fiscal.Year, 4, 7))) %>%
  ggplot(aes(x = Fiscal.Year_numeric, y = log_total_consideration, color = Municipality)) +
  geom_line() +
  geom_point() +
  labs(x = "Fiscal Year",
       y = "Natural Logarithm of Total Consideration for Taxable Conveyances",
       color = "Municipality")

# next 10 municipals
group_municipalities <- df %>%
  count(Municipality) %>%
  arrange(desc(n)) %>%
  .[121:130, "Municipality"]

# Plot with legend trimming
```

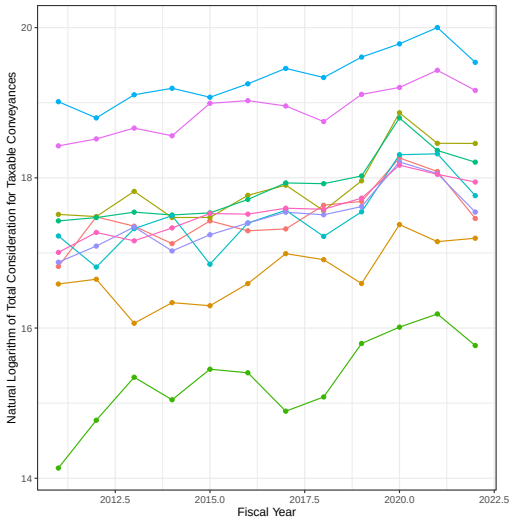
```r
plot13 <- df %>%
  filter(Municipality %in% group_municipalities) %>%
  mutate(log_total_consideration = log(Total.Consideration.for.Taxable.Conveyances),
         Fiscal.Year_numeric = as.numeric(substring(Fiscal.Year, 4, 7))) %>%
  ggplot(aes(x = Fiscal.Year_numeric, y = log_total_consideration, color = Municipality)) +
  geom_line() +
  geom_point() +
  labs(x = "Fiscal Year",
       y = "Natural Logarithm of Total Consideration for Taxable Conveyances",
       color = "Municipality")

# next 10 municipals
group_municipalities <- df %>%
  count(Municipality) %>%
  arrange(desc(n)) %>%
  .[131:140, "Municipality"]

# Plot with legend trimming
plot14 <- df %>%
  filter(Municipality %in% group_municipalities) %>%
  mutate(log_total_consideration = log(Total.Consideration.for.Taxable.Conveyances),
         Fiscal.Year_numeric = as.numeric(substring(Fiscal.Year, 4, 7))) %>%
  ggplot(aes(x = Fiscal.Year_numeric, y = log_total_consideration, color = Municipality)) +
  geom_line() +
  geom_point() +
  labs(x = "Fiscal Year",
       y = "Natural Logarithm of Total Consideration for Taxable Conveyances",
       color = "Municipality")

# next 10 municipals
group_municipalities <- df %>%
  count(Municipality) %>%
  arrange(desc(n)) %>%
  .[141:150, "Municipality"]

# Plot with legend trimming
plot15 <- df %>%
  filter(Municipality %in% group_municipalities) %>%
  mutate(log_total_consideration = log(Total.Consideration.for.Taxable.Conveyances),
         Fiscal.Year_numeric = as.numeric(substring(Fiscal.Year, 4, 7))) %>%
  ggplot(aes(x = Fiscal.Year_numeric, y = log_total_consideration, color = Municipality)) +
  geom_line() +
  geom_point() +
  labs(x = "Fiscal Year",
       y = "Natural Logarithm of Total Consideration for Taxable Conveyances",
       color = "Municipality")

# next 10 municipals
group_municipalities <- df %>%
  count(Municipality) %>%
  arrange(desc(n)) %>%
  .[151:160, "Municipality"]
```

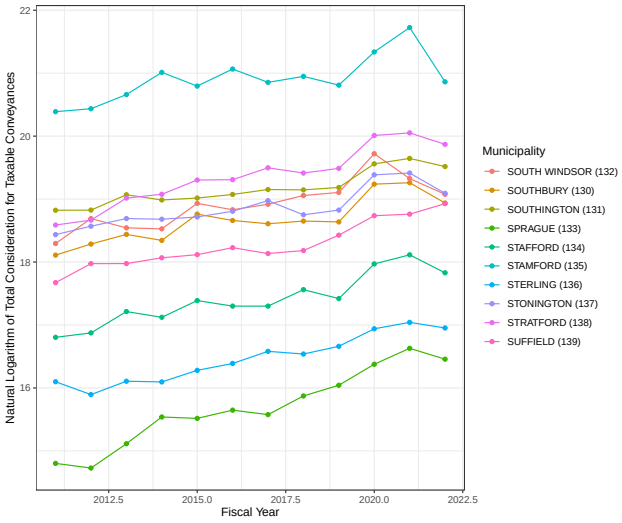
```r
# Plot with legend trimming
plot16 <- df %>%
  filter(Municipality %in% group_municipalities) %>%
  mutate(log_total_consideration = log(Total.Consideration.for.Taxable.Conveyances),
         Fiscal.Year_numeric = as.numeric(substring(Fiscal.Year, 4, 7))) %>%
  ggplot(aes(x = Fiscal.Year_numeric, y = log_total_consideration, color = Municipality)) +
  geom_line() +
  geom_point() +
  labs(x = "Fiscal Year",
       y = "Natural Logarithm of Total Consideration for Taxable Conveyances",
       color = "Municipality")

# next 10 municipals
group_municipalities <- df %>%
  count(Municipality) %>%
  arrange(desc(n)) %>%
  .[161:171, "Municipality"]

# Plot with legend trimming
plot17 <- df %>%
  filter(Municipality %in% group_municipalities) %>%
  mutate(log_total_consideration = log(Total.Consideration.for.Taxable.Conveyances),
         Fiscal.Year_numeric = as.numeric(substring(Fiscal.Year, 4, 7))) %>%
  ggplot(aes(x = Fiscal.Year_numeric, y = log_total_consideration, color = Municipality)) +
  geom_line() +
  geom_point() +
  labs(x = "Fiscal Year",
       y = "Natural Logarithm of Total Consideration for Taxable Conveyances",
       color = "Municipality")

grid.arrange(plot1, plot2,
             plot3, plot4,
             plot5, plot6,
             nrow = 3)
```

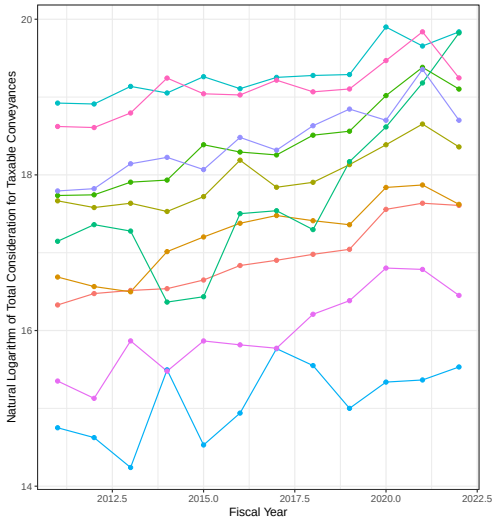

```
grid.arrange(plot7, plot8,
             plot9, plot10,
             plot11, plot12,
             nrow = 3)
```

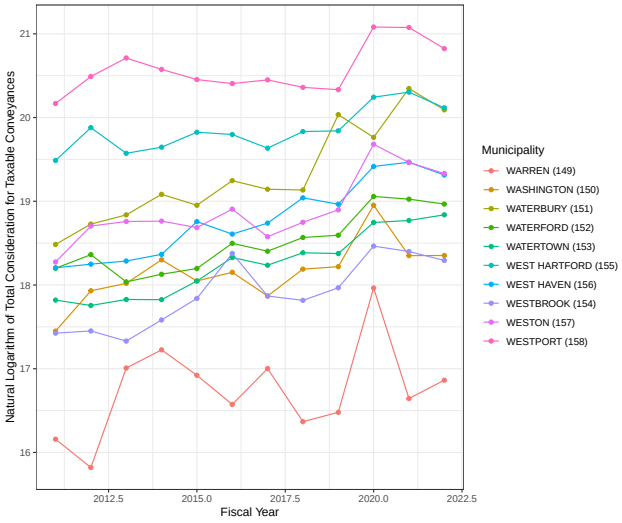


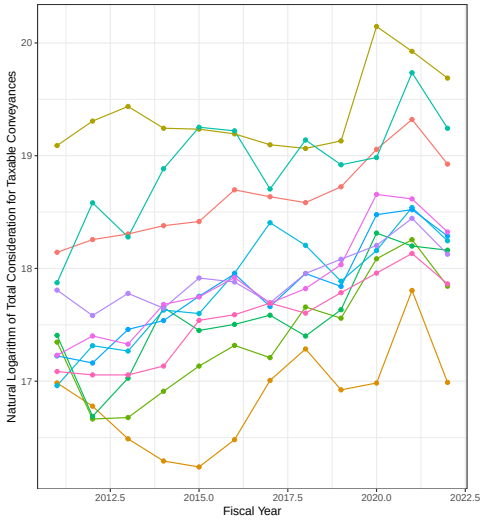

```
grid.arrange(plot13, plot14,
             plot15, plot16,
             plot17, nrow = 3)
```





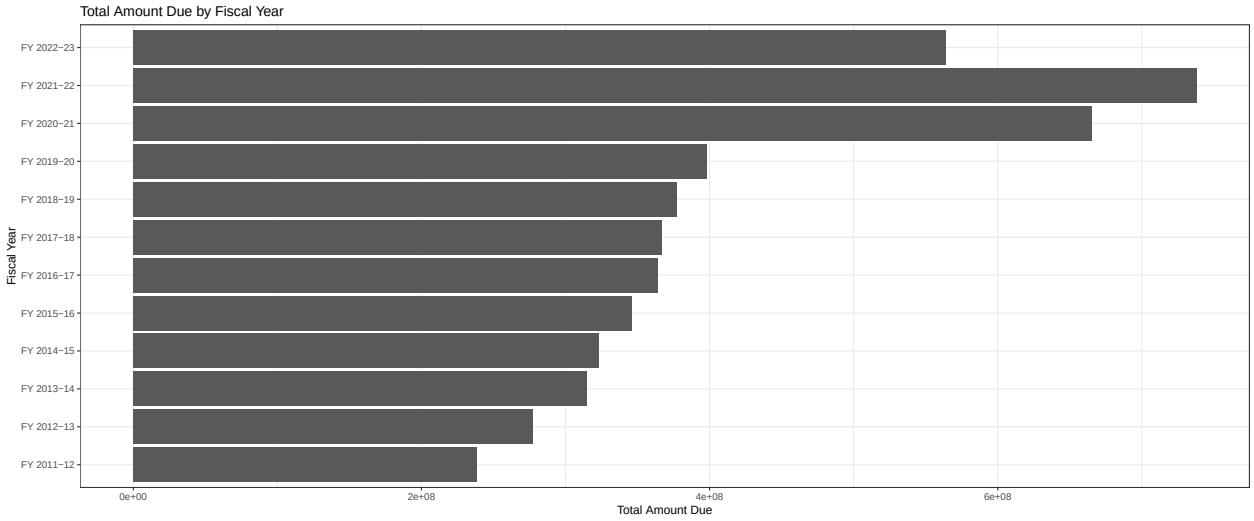


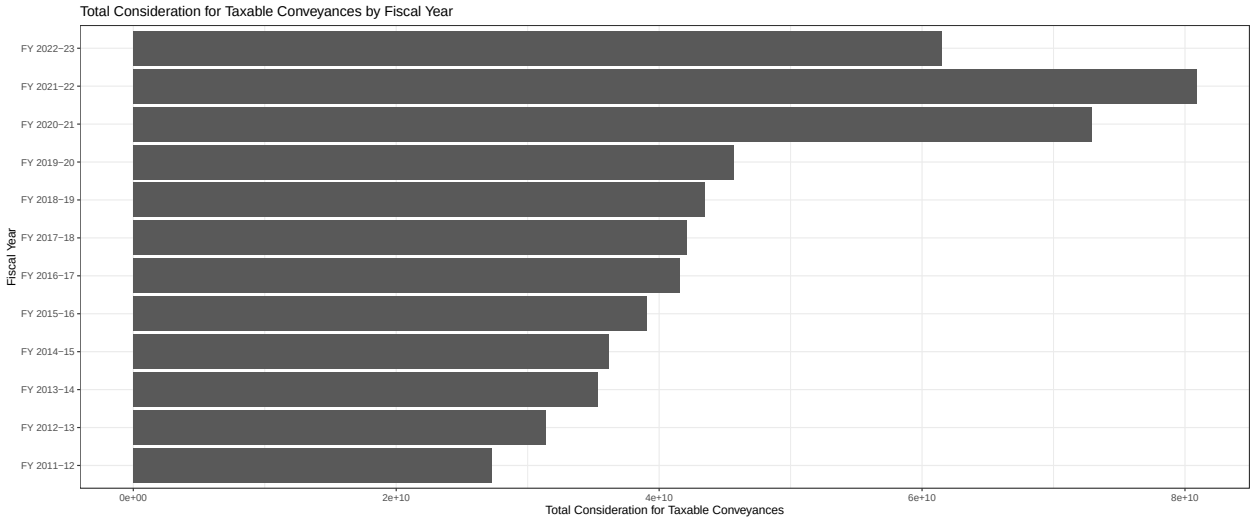


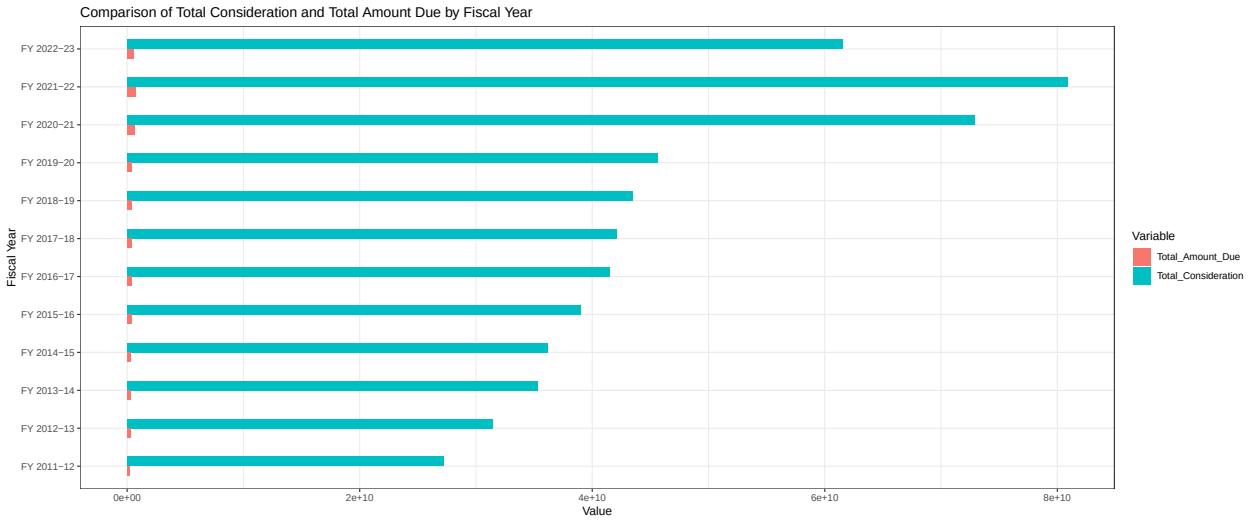



```
par(mfrow = c(3, 1))

plot_amount <- df |>
  group_by(Fiscal.Year) |>
```

```
  ggplot(aes(x = Total.Amount.Due, y = Fiscal.Year)) +
  geom_col() +
  labs(x = "Total Amount Due",
       y = "Fiscal Year",
       title = "Total Amount Due by Fiscal Year")

plot_consideration <- df |>
  group_by(Fiscal.Year) |>
  ggplot(aes(x = Total.Consideration.for.Taxable.Conveyances, y = Fiscal.Year)) +
  geom_col() +
  labs(x = "Total Consideration for Taxable Conveyances",
       y = "Fiscal Year",
       title = "Total Consideration for Taxable Conveyances by Fiscal Year")

plot_compare <- df %>%
  group_by(Fiscal.Year) %>%
  summarise(Total_Consideration = sum(Total.Consideration.for.Taxable.Conveyances),
            Total_Amount_Due = sum(Total.Amount.Due)) %>%
  pivot_longer(cols = c(Total_Consideration, Total_Amount_Due), names_to = "Variable", values_to = "Valu
  ggplot(aes(x = Value, y = Fiscal.Year, fill = Variable)) +
  geom_col(position = "dodge", width = 0.5) +
  labs(x = "Value",
       y = "Fiscal Year",
       title = "Comparison of Total Consideration and Total Amount Due by Fiscal Year",
       fill = "Variable")

grid.arrange(plot_amount,
             plot_consideration,
             plot_compare, nrow = 3)
```



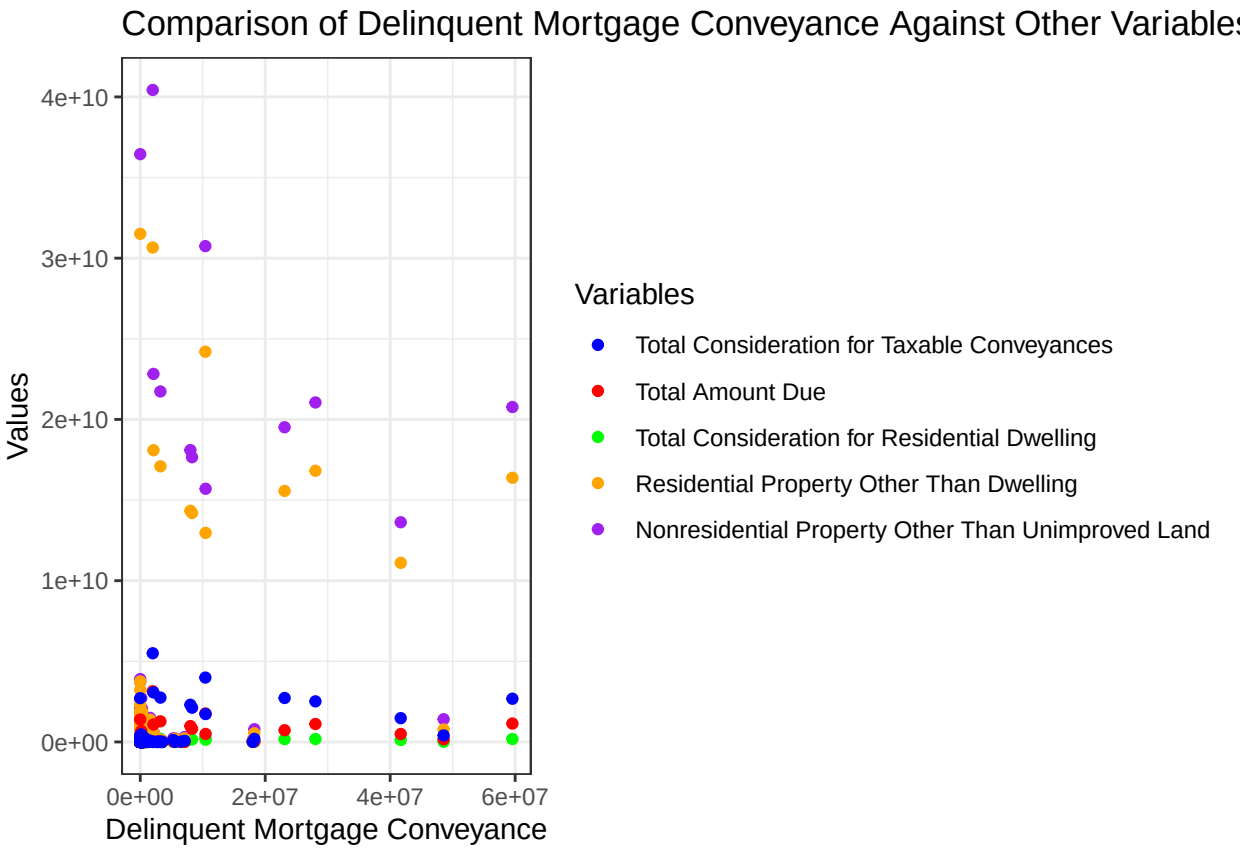




graph on deliquent mortgage for comparisons against all other variables

```
df |>
  ggplot(aes(x = Delinquent.Mortgage.Conveyance)) +
  geom_point(aes(y = Total.Consideration.for.Taxable.Conveyances, color = "Total Consideration for Taxal
```

```
  geom_point(aes(y = Total.Amount.Due, color = "Total Amount Due")) +
  geom_point(aes(y = Total.Consideration.for.Residential.Dwelling, color = "Total Consideration for Resi
  geom_point(aes(y = Residential.Property.Other.Than.Dwelling, color = "Residential Property Other Than
  geom_point(aes(y = Nonresidential.Property.Other.Than.Unimproved.Land, color = "Nonresidential Proper
  labs(x = "Delinquent Mortgage Conveyance",
       y = "Values",
       title = "Comparison of Delinquent Mortgage Conveyance Against Other Variables") +
  scale_color_manual(name = "Variables",
                     values = c("blue", "red", "green", "orange", "purple"),
                     labels = c("Total Consideration for Taxable Conveyances",
                                "Total Amount Due",
                                "Total Consideration for Residential Dwelling",
                                "Residential Property Other Than Dwelling",
                                "Nonresidential Property Other Than Unimproved Land"))
```



heat correlation map to see correlation between all variables, diagonal is going from bottom left to top right.

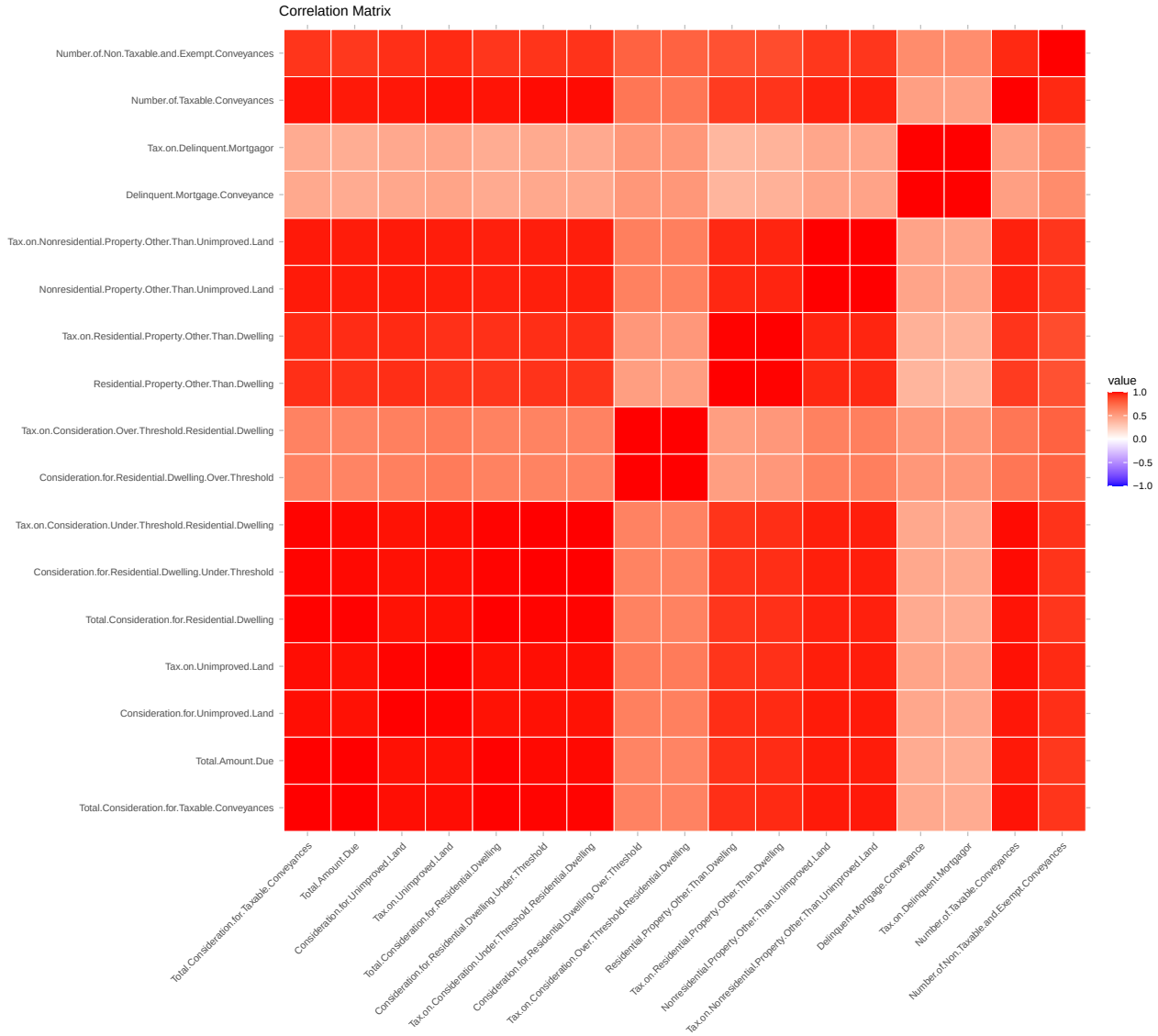
```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
# Select numeric variables
numeric_df <- df[, sapply(df, is.numeric)]
numeric_df <- numeric_df |>
  select(-Imputed)

# Calculate correlation matrix
correlation_matrix <- cor(numeric_df)
```

```
# Plot correlation matrix as a heatmap
data = reshape2::melt(correlation_matrix)
data |>
  ggplot() +
  geom_tile(aes(Var2, Var1, fill = value), color = "white") + # Add white border around tiles
  scale_fill_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0, limits = c(-1,1)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        panel.grid.major = element_line(color = "gray", size = 0.5)) + # Add gray grid lines
  labs(title = "Correlation Matrix", x = "", y = "") +
  coord_fixed()
```

```
## Warning: The `size` argument of `element_line()` is deprecated as of ggplot2 3.4.0.
## i Please use the `linewidth` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



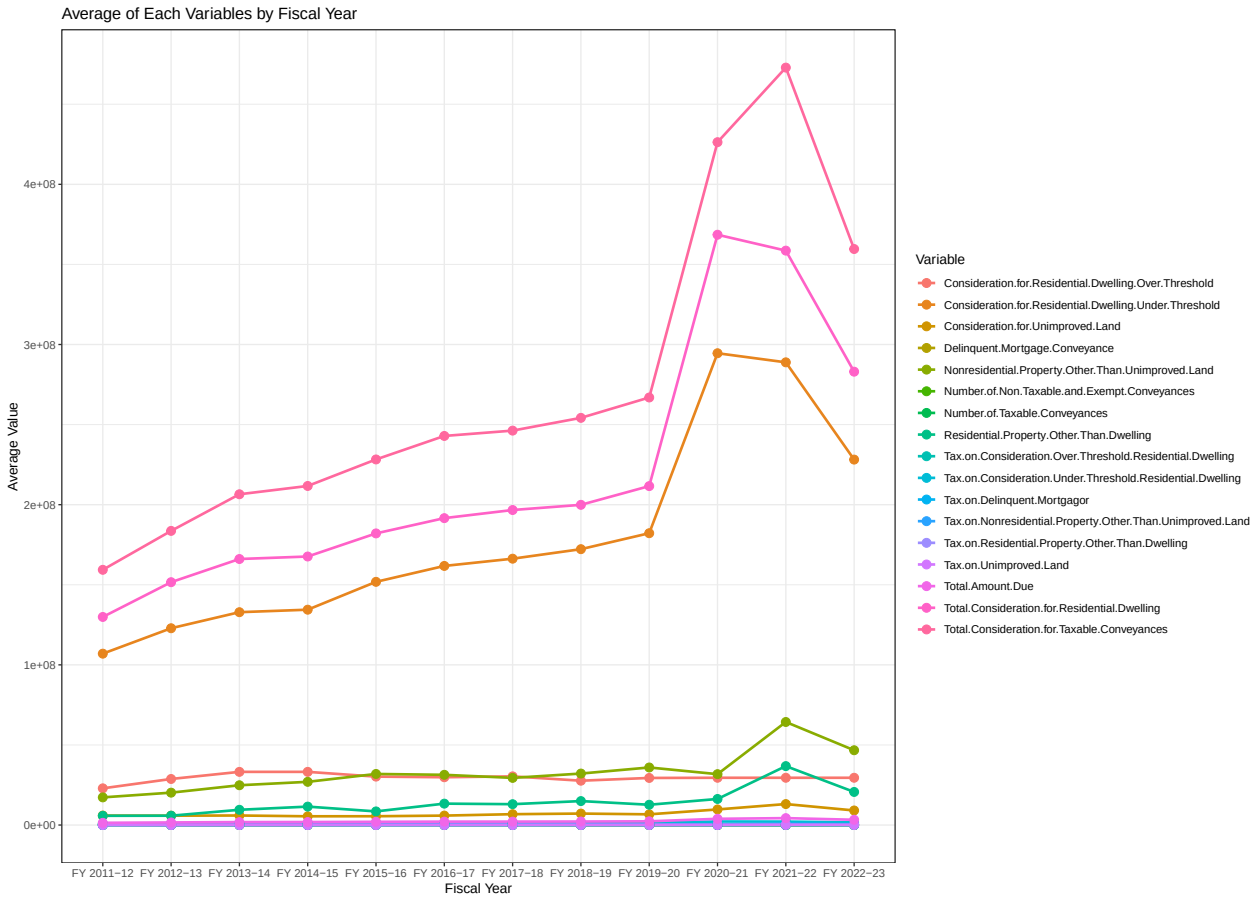
graph on all variables' values by fiscal year

```
# Select numeric variables and exclude 'Municipality' and 'Imputed' columns
numeric_df <- df %>%
  select(-Municipality, -Imputed) %>%
  select_if(is.numeric)

numeric_df$Fiscal.Year <- df$Fiscal.Year

tidy_df <- numeric_df %>%
  gather(key = "Variable", value = "Value", -Fiscal.Year)

# Calculate average by fiscal year and variable
average_df <- tidy_df %>%
  group_by(Fiscal.Year, Variable) %>%
  summarize(Average = mean(Value))
```

```
## `summarise()` has grouped output by 'Fiscal.Year'. You can override using the
## `.groups` argument.
```

```
# Plot line plot
ggplot(average_df, aes(x = Fiscal.Year, y = Average, color = Variable, group = Variable)) +
  geom_line(size=1) +
  geom_point(size=3) +
  labs(title = "Average of Each Variables by Fiscal Year",
       x = "Fiscal Year",
       y = "Average Value")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



NEW PART HERE SALISA

Data Preprocessing

Our target variable: Total.Consideration.for.Taxable.Conveyances

so we remove total.amount.due because it shows high correlation and similarity in technicality

```
model_df <- df |>
  dplyr::select(!c(Total.Amount.Due))
```

PCA as we see high correlation and redundant information.

```
pca <- model_df |>
  dplyr::select(!c(Fiscal.Year, Town.Code, Municipality, Imputed, Total.Consideration.for.Taxable.Conve
  prcomp(scale = TRUE)

pca$rotation
```

```
##                                                        PC1         PC2
## Consideration.for.Unimproved.Land                0.2827488  0.11275406
## Tax.on.Unimproved.Land                           0.2837717  0.09021224
## Total.Consideration.for.Residential.Dwelling     0.2808426  0.12326867
## Consideration.for.Residential.Dwelling.Under.Threshold    0.2818495  0.12039241
## Tax.on.Consideration.Under.Threshold.Residential.Dwelling 0.2818921  0.11869153
```

```
## Consideration.for.Residential.Dwelling.Over.Threshold     0.2112693 -0.32229031
## Tax.on.Consideration.Over.Threshold.Residential.Dwelling  0.2112691 -0.32228858
## Residential.Property.Other.Than.Dwelling                  0.2656999  0.20100171
## Tax.on.Residential.Property.Other.Than.Dwelling           0.2699126  0.18143728
## Nonresidential.Property.Other.Than.Unimproved.Land        0.2813320  0.10479170
## Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land 0.2818693  0.09927080
## Delinquent.Mortgage.Conveyance                            0.1634162 -0.55554680
## Tax.on.Delinquent.Mortgagor                               0.1613295 -0.56180461
## Number.of.Taxable.Conveyances                             0.2838545  0.06395335
## Number.of.Non.Taxable.and.Exempt.Conveyances              0.2777226 -0.07159793
##                                                                  PC3
## Consideration.for.Unimproved.Land                         0.0250114771
## Tax.on.Unimproved.Land                                    0.0006665089
## Total.Consideration.for.Residential.Dwelling              0.0133037864
## Consideration.for.Residential.Dwelling.Under.Threshold    0.0308784700
## Tax.on.Consideration.Under.Threshold.Residential.Dwelling 0.0287014562
## Consideration.for.Residential.Dwelling.Over.Threshold    -0.5610629623
## Tax.on.Consideration.Over.Threshold.Residential.Dwelling -0.5610653184
## Residential.Property.Other.Than.Dwelling                  0.1592306518
## Tax.on.Residential.Property.Other.Than.Dwelling           0.1382318308
## Nonresidential.Property.Other.Than.Unimproved.Land        0.0453282854
## Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land 0.0388906059
## Delinquent.Mortgage.Conveyance                            0.3996495617
## Tax.on.Delinquent.Mortgagor                               0.3937435972
## Number.of.Taxable.Conveyances                            -0.0121040965
## Number.of.Non.Taxable.and.Exempt.Conveyances             -0.0700280265
##                                                                  PC4
## Consideration.for.Unimproved.Land                        -0.10719624
## Tax.on.Unimproved.Land                                   -0.20368942
## Total.Consideration.for.Residential.Dwelling             -0.25279261
## Consideration.for.Residential.Dwelling.Under.Threshold   -0.24585604
## Tax.on.Consideration.Under.Threshold.Residential.Dwelling -0.24987073
## Consideration.for.Residential.Dwelling.Over.Threshold     0.15600162
## Tax.on.Consideration.Over.Threshold.Residential.Dwelling  0.15600673
## Residential.Property.Other.Than.Dwelling                  0.54134286
## Tax.on.Residential.Property.Other.Than.Dwelling           0.48139496
## Nonresidential.Property.Other.Than.Unimproved.Land        0.17881232
## Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land 0.16652626
## Delinquent.Mortgage.Conveyance                            0.01050532
## Tax.on.Delinquent.Mortgagor                               0.01008249
## Number.of.Taxable.Conveyances                            -0.25485248
## Number.of.Non.Taxable.and.Exempt.Conveyances             -0.24879802
##                                                                  PC5
## Consideration.for.Unimproved.Land                         0.08670995
## Tax.on.Unimproved.Land                                    0.05881594
## Total.Consideration.for.Residential.Dwelling             -0.33116148
## Consideration.for.Residential.Dwelling.Under.Threshold   -0.24939877
## Tax.on.Consideration.Under.Threshold.Residential.Dwelling -0.25393592
## Consideration.for.Residential.Dwelling.Over.Threshold    -0.10220194
## Tax.on.Consideration.Over.Threshold.Residential.Dwelling -0.10221277
## Residential.Property.Other.Than.Dwelling                  0.01952735
## Tax.on.Residential.Property.Other.Than.Dwelling          -0.02030199
## Nonresidential.Property.Other.Than.Unimproved.Land        0.09297320
## Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land 0.08347571
```

```
## Delinquent.Mortgage.Conveyance                                -0.05260428
## Tax.on.Delinquent.Mortgagor                                   -0.06789553
## Number.of.Taxable.Conveyances                                 -0.08930061
## Number.of.Non.Taxable.and.Exempt.Conveyances                   0.83667255
##                                                                       PC6
## Consideration.for.Unimproved.Land                              0.08171724
## Tax.on.Unimproved.Land                                         0.08392563
## Total.Consideration.for.Residential.Dwelling                   0.03589693
## Consideration.for.Residential.Dwelling.Under.Threshold         0.04679279
## Tax.on.Consideration.Under.Threshold.Residential.Dwelling      0.05357535
## Consideration.for.Residential.Dwelling.Over.Threshold          0.03251606
## Tax.on.Consideration.Over.Threshold.Residential.Dwelling       0.03250873
## Residential.Property.Other.Than.Dwelling                       0.29618080
## Tax.on.Residential.Property.Other.Than.Dwelling                0.33951590
## Nonresidential.Property.Other.Than.Unimproved.Land            -0.61062018
## Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land     -0.60328064
## Delinquent.Mortgage.Conveyance                                -0.01149228
## Tax.on.Delinquent.Mortgagor                                    0.01498082
## Number.of.Taxable.Conveyances                                  0.08073899
## Number.of.Non.Taxable.and.Exempt.Conveyances                   0.17647171
##                                                                       PC7
## Consideration.for.Unimproved.Land                             -0.63949389
## Tax.on.Unimproved.Land                                        -0.48509306
## Total.Consideration.for.Residential.Dwelling                  -0.07380212
## Consideration.for.Residential.Dwelling.Under.Threshold         0.18014930
## Tax.on.Consideration.Under.Threshold.Residential.Dwelling      0.18852924
## Consideration.for.Residential.Dwelling.Over.Threshold         -0.01828341
## Tax.on.Consideration.Over.Threshold.Residential.Dwelling      -0.01826131
## Residential.Property.Other.Than.Dwelling                      -0.01401922
## Tax.on.Residential.Property.Other.Than.Dwelling                0.15744242
## Nonresidential.Property.Other.Than.Unimproved.Land             0.02486899
## Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land      0.05170536
## Delinquent.Mortgage.Conveyance                                -0.02857349
## Tax.on.Delinquent.Mortgagor                                   -0.01167531
## Number.of.Taxable.Conveyances                                  0.45197417
## Number.of.Non.Taxable.and.Exempt.Conveyances                   0.21934164
##                                                                       PC8
## Consideration.for.Unimproved.Land                             -0.207561777
## Tax.on.Unimproved.Land                                        -0.247130224
## Total.Consideration.for.Residential.Dwelling                   0.622847272
## Consideration.for.Residential.Dwelling.Under.Threshold         0.093197493
## Tax.on.Consideration.Under.Threshold.Residential.Dwelling      0.076835089
## Consideration.for.Residential.Dwelling.Over.Threshold          0.004467568
## Tax.on.Consideration.Over.Threshold.Residential.Dwelling       0.004476690
## Residential.Property.Other.Than.Dwelling                       0.069422806
## Tax.on.Residential.Property.Other.Than.Dwelling               -0.021107485
## Nonresidential.Property.Other.Than.Unimproved.Land             0.020719025
## Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land     -0.019785630
## Delinquent.Mortgage.Conveyance                                 0.039685483
## Tax.on.Delinquent.Mortgagor                                   -0.041498968
## Number.of.Taxable.Conveyances                                 -0.642901557
## Number.of.Non.Taxable.and.Exempt.Conveyances                   0.265652629
##                                                                       PC9
## Consideration.for.Unimproved.Land                              0.444967417
```

```
## Tax.on.Unimproved.Land                                          -0.580967749
## Total.Consideration.for.Residential.Dwelling                    -0.032185982
## Consideration.for.Residential.Dwelling.Under.Threshold           0.057480197
## Tax.on.Consideration.Under.Threshold.Residential.Dwelling        0.068598644
## Consideration.for.Residential.Dwelling.Over.Threshold            0.008745671
## Tax.on.Consideration.Over.Threshold.Residential.Dwelling         0.008749105
## Residential.Property.Other.Than.Dwelling                         0.426955034
## Tax.on.Residential.Property.Other.Than.Dwelling                 -0.489587029
## Nonresidential.Property.Other.Than.Unimproved.Land              -0.004267356
## Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land       -0.045902753
## Delinquent.Mortgage.Conveyance                                  -0.084966828
## Tax.on.Delinquent.Mortgagor                                      0.087234510
## Number.of.Taxable.Conveyances                                    0.126179697
## Number.of.Non.Taxable.and.Exempt.Conveyances                     0.020491065
##                                                                         PC10
## Consideration.for.Unimproved.Land                               -0.003901609
## Tax.on.Unimproved.Land                                           0.044727852
## Total.Consideration.for.Residential.Dwelling                    -0.570908624
## Consideration.for.Residential.Dwelling.Under.Threshold           0.488573412
## Tax.on.Consideration.Under.Threshold.Residential.Dwelling        0.450133049
## Consideration.for.Residential.Dwelling.Over.Threshold            0.020564895
## Tax.on.Consideration.Over.Threshold.Residential.Dwelling         0.020549469
## Residential.Property.Other.Than.Dwelling                         0.019316173
## Tax.on.Residential.Property.Other.Than.Dwelling                 -0.039765447
## Nonresidential.Property.Other.Than.Unimproved.Land               0.033287931
## Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land       -0.053677076
## Delinquent.Mortgage.Conveyance                                   0.164646786
## Tax.on.Delinquent.Mortgagor                                     -0.163137451
## Number.of.Taxable.Conveyances                                   -0.412396375
## Number.of.Non.Taxable.and.Exempt.Conveyances                     0.011904661
##                                                                         PC11
## Consideration.for.Unimproved.Land                                0.271840567
## Tax.on.Unimproved.Land                                          -0.254000979
## Total.Consideration.for.Residential.Dwelling                    -0.085411480
## Consideration.for.Residential.Dwelling.Under.Threshold           0.061795529
## Tax.on.Consideration.Under.Threshold.Residential.Dwelling        0.136755526
## Consideration.for.Residential.Dwelling.Over.Threshold           -0.002424326
## Tax.on.Consideration.Over.Threshold.Residential.Dwelling        -0.002426415
## Residential.Property.Other.Than.Dwelling                        -0.388670155
## Tax.on.Residential.Property.Other.Than.Dwelling                  0.365485755
## Nonresidential.Property.Other.Than.Unimproved.Land              -0.086348081
## Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land        0.122069766
## Delinquent.Mortgage.Conveyance                                  -0.500241465
## Tax.on.Delinquent.Mortgagor                                      0.497791617
## Number.of.Taxable.Conveyances                                   -0.161942027
## Number.of.Non.Taxable.and.Exempt.Conveyances                     0.022225628
##                                                                         PC12
## Consideration.for.Unimproved.Land                               -0.3960801884
## Tax.on.Unimproved.Land                                           0.3982913622
## Total.Consideration.for.Residential.Dwelling                    -0.0564350181
## Consideration.for.Residential.Dwelling.Under.Threshold           0.0455120954
## Tax.on.Consideration.Under.Threshold.Residential.Dwelling        0.0571975078
## Consideration.for.Residential.Dwelling.Over.Threshold            0.0009163820
## Tax.on.Consideration.Over.Threshold.Residential.Dwelling         0.0008839086
```

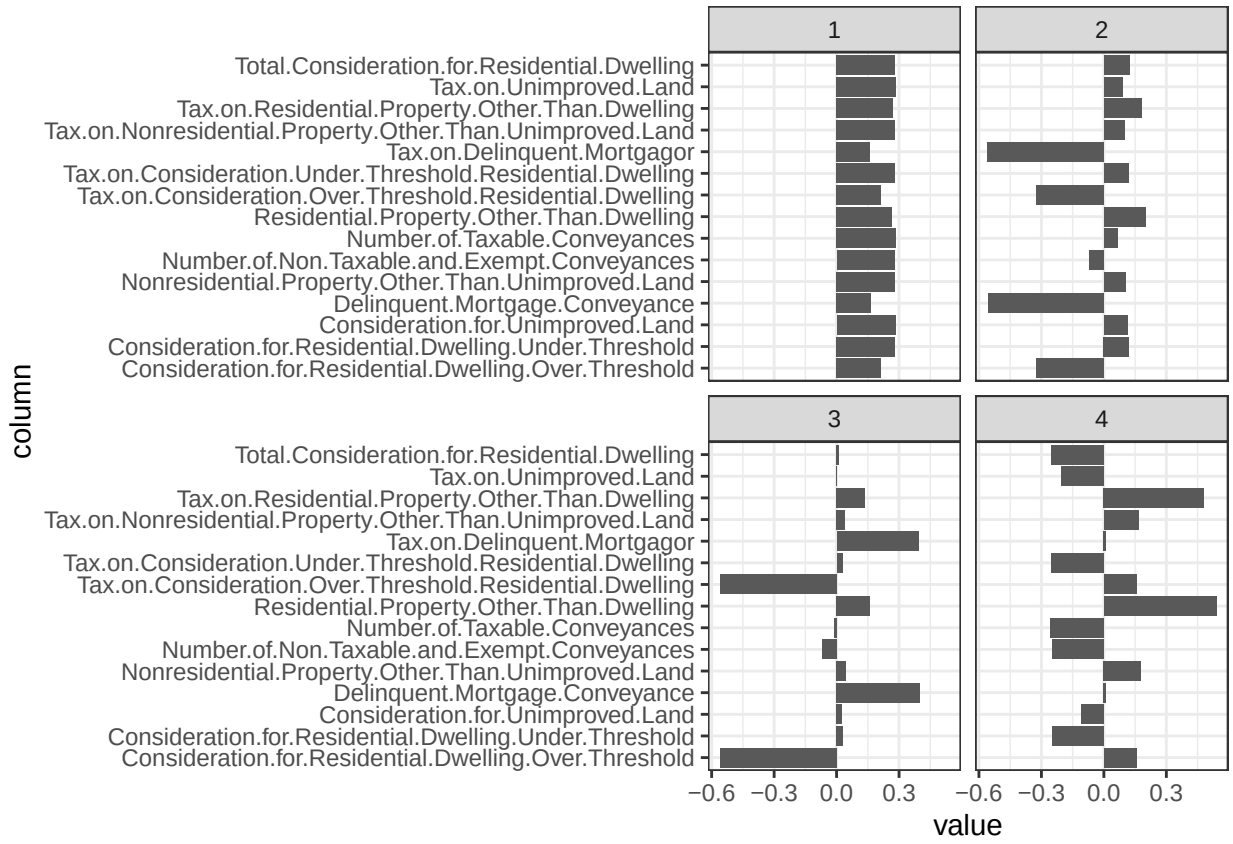
```
## Residential.Property.Other.Than.Dwelling                     0.3735405986
## Tax.on.Residential.Property.Other.Than.Dwelling             -0.3524186586
## Nonresidential.Property.Other.Than.Unimproved.Land          -0.0286668997
## Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land    0.0328789712
## Delinquent.Mortgage.Conveyance                              -0.4504207701
## Tax.on.Delinquent.Mortgagor                                  0.4498251883
## Number.of.Taxable.Conveyances                               -0.0712278185
## Number.of.Non.Taxable.and.Exempt.Conveyances                 0.0050751543
##                                                                     PC13
## Consideration.for.Unimproved.Land                            0.006387410
## Tax.on.Unimproved.Land                                      -0.008882966
## Total.Consideration.for.Residential.Dwelling                 0.006698269
## Consideration.for.Residential.Dwelling.Under.Threshold       0.100606061
## Tax.on.Consideration.Under.Threshold.Residential.Dwelling -0.120654359
## Consideration.for.Residential.Dwelling.Over.Threshold        0.001215244
## Tax.on.Consideration.Over.Threshold.Residential.Dwelling     0.001122076
## Residential.Property.Other.Than.Dwelling                    -0.064619146
## Tax.on.Residential.Property.Other.Than.Dwelling              0.058076680
## Nonresidential.Property.Other.Than.Unimproved.Land           0.685419694
## Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land -0.688139807
## Delinquent.Mortgage.Conveyance                              -0.108282725
## Tax.on.Delinquent.Mortgagor                                  0.108564306
## Number.of.Taxable.Conveyances                                0.027305478
## Number.of.Non.Taxable.and.Exempt.Conveyances                -0.003406635
##                                                                     PC14
## Consideration.for.Unimproved.Land                           -0.0127884010
## Tax.on.Unimproved.Land                                       0.0140911501
## Total.Consideration.for.Residential.Dwelling                -0.0034207623
## Consideration.for.Residential.Dwelling.Under.Threshold      -0.6936277234
## Tax.on.Consideration.Under.Threshold.Residential.Dwelling    0.7003804109
## Consideration.for.Residential.Dwelling.Over.Threshold       -0.0002628268
## Tax.on.Consideration.Over.Threshold.Residential.Dwelling    -0.0004364776
## Residential.Property.Other.Than.Dwelling                     0.0062059923
## Tax.on.Residential.Property.Other.Than.Dwelling             -0.0057694490
## Nonresidential.Property.Other.Than.Unimproved.Land           0.1186992805
## Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land -0.1138602192
## Delinquent.Mortgage.Conveyance                               0.0192000085
## Tax.on.Delinquent.Mortgagor                                 -0.0192584615
## Number.of.Taxable.Conveyances                               -0.0103693708
## Number.of.Non.Taxable.and.Exempt.Conveyances                 0.0011288514
##                                                                     PC15
## Consideration.for.Unimproved.Land                           -1.774358e-06
## Tax.on.Unimproved.Land                                      -2.253003e-05
## Total.Consideration.for.Residential.Dwelling                 1.195326e-05
## Consideration.for.Residential.Dwelling.Under.Threshold       7.648144e-05
## Tax.on.Consideration.Under.Threshold.Residential.Dwelling -8.003246e-05
## Consideration.for.Residential.Dwelling.Over.Threshold        7.071080e-01
## Tax.on.Consideration.Over.Threshold.Residential.Dwelling    -7.071056e-01
## Residential.Property.Other.Than.Dwelling                    -3.229922e-06
## Tax.on.Residential.Property.Other.Than.Dwelling              6.102141e-06
## Nonresidential.Property.Other.Than.Unimproved.Land          -5.567170e-05
## Tax.on.Nonresidential.Property.Other.Than.Unimproved.Land    6.265287e-05
## Delinquent.Mortgage.Conveyance                               1.280000e-05
## Tax.on.Delinquent.Mortgagor                                 -1.519450e-05
```

```
## Number.of.Taxable.Conveyances                                 8.486510e-06
## Number.of.Non.Taxable.and.Exempt.Conveyances                 -3.223927e-06
```

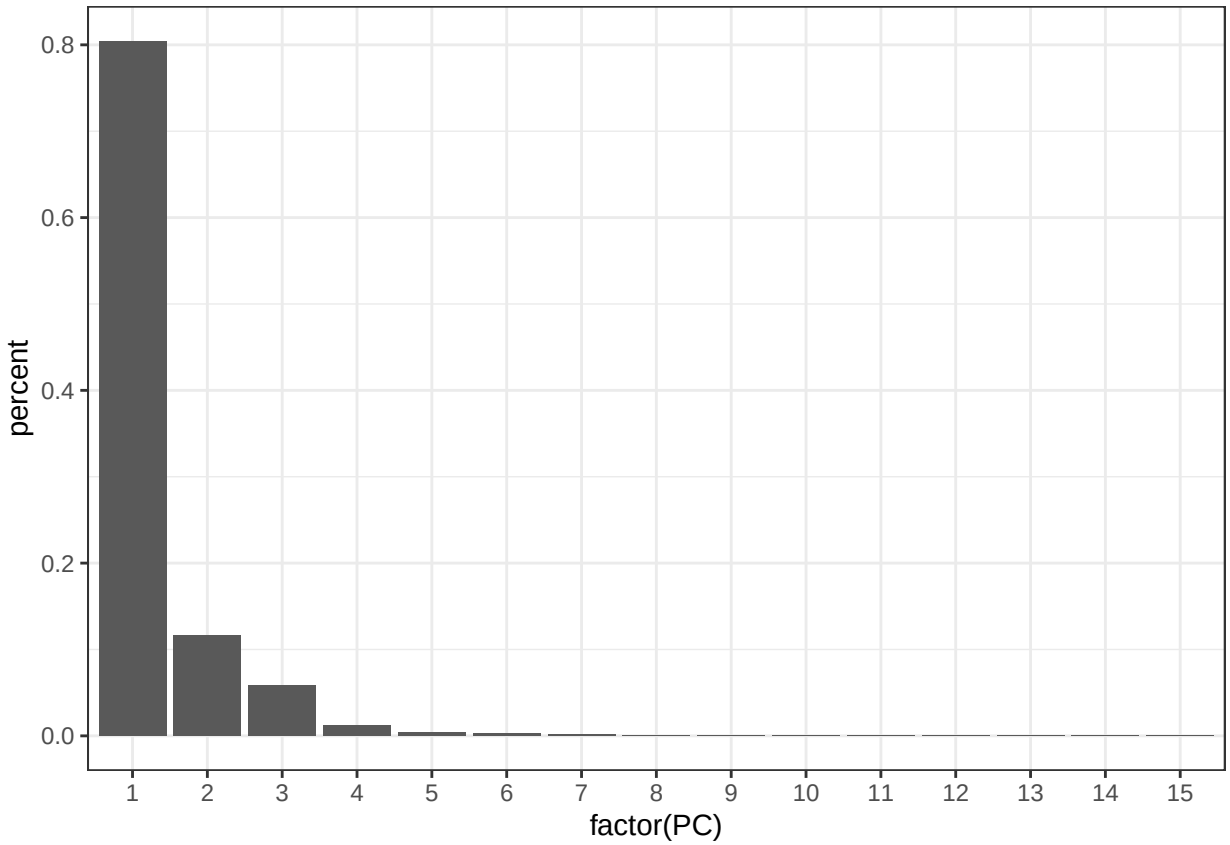
```
library(broom)
pca |>
  tidy(matrix = "pcs")
```

```
## # A tibble: 15 x 4
##       PC    std.dev percent cumulative
##    <dbl>      <dbl>   <dbl>      <dbl>
##  1     1 3.47       0.804        0.804
##  2     2 1.32       0.116        0.921
##  3     3 0.936      0.0584       0.979
##  4     4 0.417      0.0116       0.990
##  5     5 0.245      0.00401      0.994
##  6     6 0.219      0.00321      0.998
##  7     7 0.149      0.00148      0.999
##  8     8 0.0947     0.0006       1.00
##  9     9 0.0414     0.00011      1.00
## 10    10 0.0306     0.00006      1.00
## 11    11 0.0155     0.00002      1.00
## 12    12 0.0145     0.00001      1
## 13    13 0.00393    0            1
## 14    14 0.000913   0            1
## 15    15 0.00000844 0            1
```

```
model_loadings <- pca |>
  tidy(matrix = "loadings")

model_variances <- pca |>
  tidy(matrix = "pcs")

model_loadings |>
  filter(PC <=4) |>
  ggplot(aes(y = column,
             x = value)) +
  geom_col(aes(y = column)
           ) +
  facet_wrap(~PC)
```




```
model_variances |>
  ggplot(aes(x = factor(PC),
             y = percent)) +
  geom_col()
```



WE WILL CHOOSE PC1, PC2, PC3 and PC 4 which in total explain 99.049% of variability on a bases across different models. For the same model, like linear regression for example, we may test different PC values like one with 4 and another with 6.

Modeling

data splitting 75% training 25% test

```
library(tidymodels)
```

```
## -- Attaching packages -------------------------------------- tidymodels 1.2.0 --
## v dials        1.2.1      v rsample      1.2.1
## v infer        1.0.7      v tune         1.2.0
## v modeldata    1.3.0      v workflows    1.1.4
## v parsnip      1.2.1      v workflowsets 1.1.0
## v recipes      1.0.10     v yardstick    1.3.1
## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x gridExtra::combine() masks dplyr::combine()
## x scales::discard()    masks purrr::discard()
## x dplyr::filter()      masks stats::filter()
## x recipes::fixed()     masks stringr::fixed()
## x dplyr::lag()         masks stats::lag()
## x yardstick::spec()    masks readr::spec()
## x recipes::step()      masks stats::step()
## * Dig deeper into tidy modeling with R at https://www.tmwr.org
```

```r
set.seed(1)

# remove imputed column

new_df <- model_df |>
  dplyr::select(!c(Imputed))

model_split <- new_df |>
  initial_split(prop = 0.75,
                strata = Total.Consideration.for.Taxable.Conveyances)

model_split
```

```
## <Training/Testing/Total>
## <1536/516/2052>
```

```r
model_train <- model_split |>
  training()

model_test <- model_split |>
  testing()

model_train |>
  count()
```

```
##      n
## 1 1536
```

```r
model_test |>
  count()
```

```
##     n
## 1 516
```

Model 1: Linear Regression with basic 4 PC

```r
consideration_parsnip_1 <- linear_reg() |>
  set_mode("regression") |>
  set_engine("lm")

consideration_recipe_1 <- recipe(Total.Consideration.for.Taxable.Conveyances ~ .,
                      data = model_train)

consideration_recipe_1 <- consideration_recipe_1 |>
  step_normalize(all_numeric_predictors()) |>
  step_pca(all_numeric_predictors(),
           num_comp = 4) |>
  step_dummy(all_nominal_predictors())
```

```r
consideration_workflow_1 <- workflow() |>
  add_model(consideration_parsnip_1) |>
  add_recipe(consideration_recipe_1)
```

Model 2: Linear Regression with 6 PC

```
consideration_recipe_2 <- recipe(Total.Consideration.for.Taxable.Conveyances ~ .,
                        data = model_train)

consideration_recipe_2 <- consideration_recipe_2 |>
  step_normalize(all_numeric_predictors()) |>
  step_pca(all_numeric_predictors(),
           num_comp = 6) |>
  step_dummy(all_nominal_predictors())
```

```
consideration_workflow_2 <- workflow() |>
  add_model(consideration_parsnip_1) |>
  add_recipe(consideration_recipe_2)
```

Model 3, Regular Linear regression with no pca for comparison

```
consideration_workflow_3 <- workflow() |>
  add_model(consideration_parsnip_1) |>
  add_formula(Total.Consideration.for.Taxable.Conveyances ~ .)
```

Model 4: KNN with 4 PCA

```
library(kknn)
consideration_parsnip_2 <- nearest_neighbor() |>
  set_mode("regression") |>
  set_engine("kknn",
             neighbors = 5)

consideration_recipe_4 <- recipe(Total.Consideration.for.Taxable.Conveyances ~ .,
                        data = model_train)

consideration_recipe_4 <- consideration_recipe_4 |>
  step_normalize(all_numeric_predictors()) |>
  step_pca(all_numeric_predictors(),
           num_comp = 4) |>
  step_dummy(all_nominal_predictors())

consideration_workflow_4 <- workflow() |>
  add_model(consideration_parsnip_2) |>
  add_recipe(consideration_recipe_4)
```

Model 5: Knn on all for comparison

```
library(kknn)

consideration_workflow_5 <- workflow() |>
  add_model(consideration_parsnip_2) |>
  add_formula(Total.Consideration.for.Taxable.Conveyances ~ .)
```

Model 6: Random forest with 4 PCA

```
library(ranger)
consideration_parsnip_3 <- rand_forest() |>
  set_mode("regression") |>
  set_engine("ranger")

consideration_recipe_6 <- recipe(Total.Consideration.for.Taxable.Conveyances ~ .,
                       data = model_train)

consideration_recipe_6 <- consideration_recipe_6 |>
  step_normalize(all_numeric_predictors()) |>
  step_pca(all_numeric_predictors(),
           num_comp = 4) |>
  step_dummy(all_nominal_predictors())
```

```
consideration_workflow_6 <- workflow() |>
  add_model(consideration_parsnip_3) |>
  add_recipe(consideration_recipe_6)
```

Model 7: Random forest with all

```
consideration_workflow_7 <- workflow() |>
  add_model(consideration_parsnip_3) |>
  add_formula(Total.Consideration.for.Taxable.Conveyances ~ .)
```

Tibble of workflows

```
workflow_names <- c("lm_4PC",
                    "lm_6PC",
                    "lm",
                    "knn_4PC",
                    "knn",
                    "rf_4PC",
                    "rf")

workflow_objects <- list(consideration_workflow_1,
                         consideration_workflow_2,
                         consideration_workflow_3,
                         consideration_workflow_4,
                         consideration_workflow_5,
                         consideration_workflow_6,
                         consideration_workflow_7)

workflows_tbl <- tibble(work_names = workflow_names,
                        work_objects = workflow_objects)

workflows_tbl
```

```
## # A tibble: 7 x 2
##   work_names work_objects
##   <chr>      <list>
```

```
## 1 lm_4PC     <workflow>
## 2 lm_6PC     <workflow>
## 3 lm         <workflow>
## 4 knn_4PC    <workflow>
## 5 knn        <workflow>
## 6 rf_4PC     <workflow>
## 7 rf         <workflow>
```

fitting

```
set.seed(1)
workflows_tbl <- workflows_tbl |>
  rowwise() |>
  mutate(fits = list(fit(work_objects,
                         model_train)))

workflows_tbl
```

```
## # A tibble: 7 x 3
## # Rowwise:
##   work_names work_objects fits
##   <chr>      <list>       <list>
## 1 lm_4PC     <workflow>   <workflow>
## 2 lm_6PC     <workflow>   <workflow>
## 3 lm         <workflow>   <workflow>
## 4 knn_4PC    <workflow>   <workflow>
## 5 knn        <workflow>   <workflow>
## 6 rf_4PC     <workflow>   <workflow>
## 7 rf         <workflow>   <workflow>
```

```
workflows_tbl <- workflows_tbl |>
  mutate(predictions = list(predict(fits,
                                    model_test)))
```

```
## Warning: There were 3 warnings in `mutate()`.
## The first warning was:
## i In argument: `predictions = list(predict(fits, model_test))`.
## i In row 1.
## Caused by warning in `predict.lm()`:
## ! prediction from a rank-deficient fit may be misleading
## i Run `dplyr::last_dplyr_warnings()` to see the 2 remaining warnings.
```

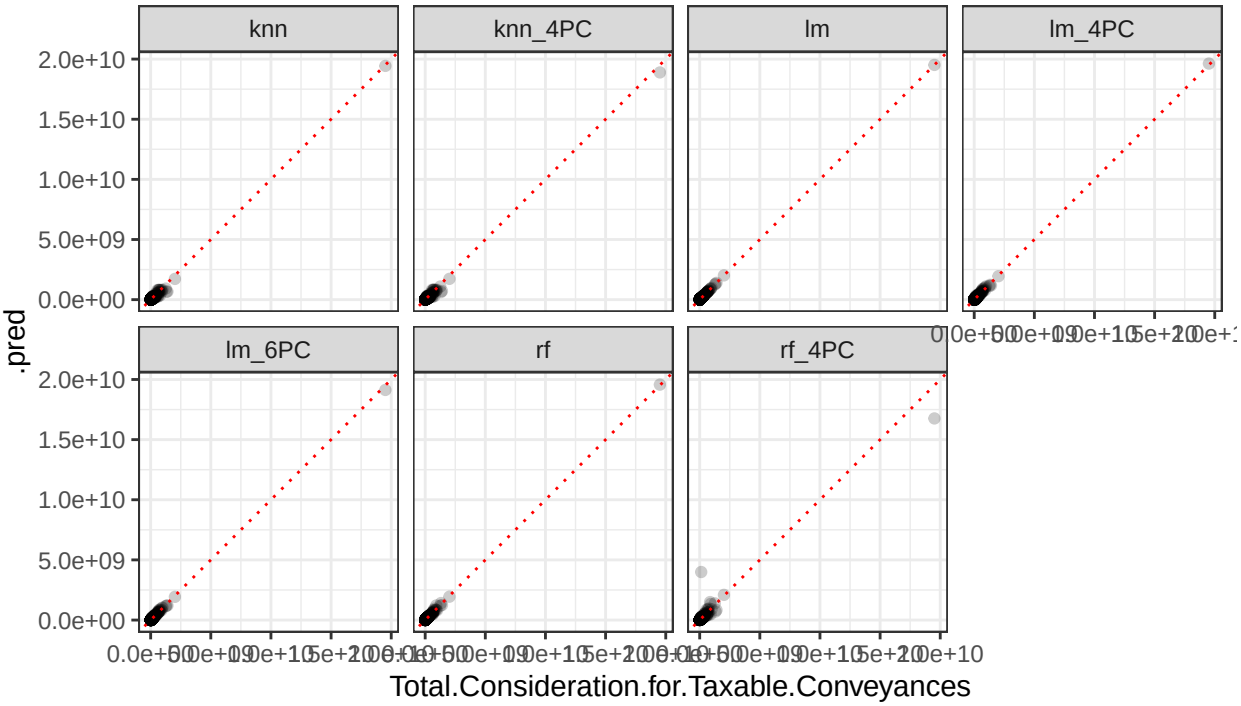
```
workflows_tbl
```

```
## # A tibble: 7 x 4
## # Rowwise:
##   work_names work_objects fits       predictions
##   <chr>      <list>       <list>     <list>
## 1 lm_4PC     <workflow>   <workflow> <tibble [516 x 1]>
## 2 lm_6PC     <workflow>   <workflow> <tibble [516 x 1]>
## 3 lm         <workflow>   <workflow> <tibble [516 x 1]>
## 4 knn_4PC    <workflow>   <workflow> <tibble [516 x 1]>
## 5 knn        <workflow>   <workflow> <tibble [516 x 1]>
## 6 rf_4PC     <workflow>   <workflow> <tibble [516 x 1]>
## 7 rf         <workflow>   <workflow> <tibble [516 x 1]>
```

plotting versus truth

```
predictions_tbl  <- workflows_tbl |>
  select(work_names,
         predictions) |>
  unnest(cols = c(predictions))

predictions_tbl <- predictions_tbl |>
  cbind(Total.Consideration.for.Taxable.Conveyances = model_test |>
          pull(Total.Consideration.for.Taxable.Conveyances))
```

```
predictions_tbl |>
 ggplot(aes(x = Total.Consideration.for.Taxable.Conveyances,
            y = .pred)) +
  geom_point(alpha = 0.2) +
  facet_wrap(~work_names, nrow = 2) +
  geom_abline(slope = 1, linetype = "dotted", color = "red") +
  coord_obs_pred() # a special coordinate function from the tidymodels family
```



metric table

```
consideration_metrics <- metric_set(yardstick::rmse,
                                    yardstick::rsq_trad,
                                    yardstick::mae)
predictions_tbl |>
  group_by(work_names) |>
  consideration_metrics(truth = Total.Consideration.for.Taxable.Conveyances,
              estimate = .pred)
```

```
## # A tibble: 21 x 4
```

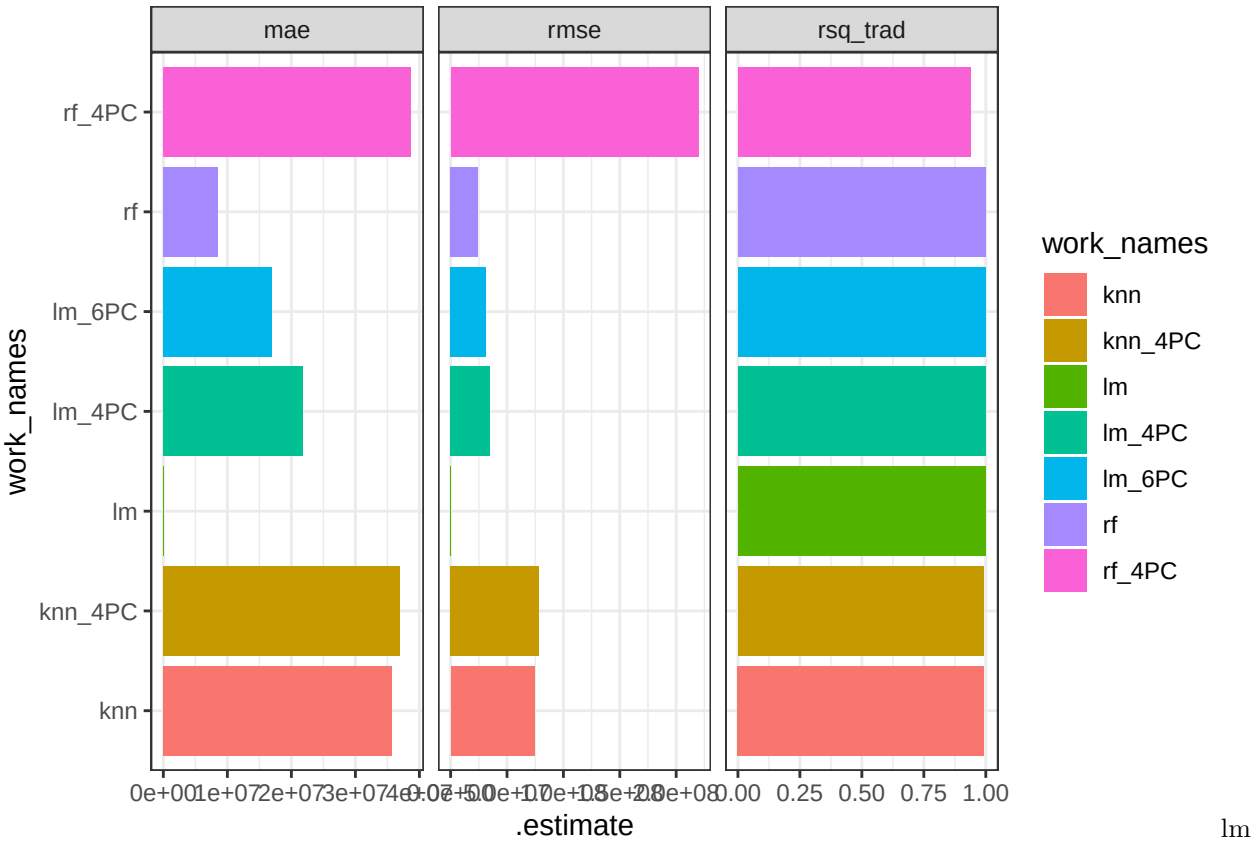
```
##    work_names .metric  .estimator     .estimate
##    <chr>      <chr>    <chr>              <dbl>
##  1 knn        rmse     standard    74293632.
##  2 knn_4PC    rmse     standard    78645801.
##  3 lm         rmse     standard           0.259
##  4 lm_4PC     rmse     standard    35196998.
##  5 lm_6PC     rmse     standard    31205262.
##  6 rf         rmse     standard    24574711.
##  7 rf_4PC     rmse     standard   219818636.
##  8 knn        rsq_trad standard           0.993
##  9 knn_4PC    rsq_trad standard           0.992
## 10 lm         rsq_trad standard           1
## # i 11 more rows
```

all combined metrics

```
predictions_metrics <- predictions_tbl |>
  group_by(work_names) |>
  consideration_metrics(truth = Total.Consideration.for.Taxable.Conveyances, estimate = .pred)

predictions_metrics
```

```
## # A tibble: 21 x 4
##    work_names .metric  .estimator     .estimate
##    <chr>      <chr>    <chr>              <dbl>
##  1 knn        rmse     standard    74293632.
##  2 knn_4PC    rmse     standard    78645801.
##  3 lm         rmse     standard           0.259
##  4 lm_4PC     rmse     standard    35196998.
##  5 lm_6PC     rmse     standard    31205262.
##  6 rf         rmse     standard    24574711.
##  7 rf_4PC     rmse     standard   219818636.
##  8 knn        rsq_trad standard           0.993
##  9 knn_4PC    rsq_trad standard           0.992
## 10 lm         rsq_trad standard           1
## # i 11 more rows
```

```
predictions_metrics  |>
  ggplot(aes(y = work_names,
             x = .estimate,
             fill = work_names)) +
  geom_col() +
  facet_wrap(~.metric,
             scales = "free_x")
```



lm

nott missing, just super small

uncertainty quantification: bootstrap

```
set.seed(1)

pc_workflows_tbl <- workflows_tbl |>
  filter(grepl("PC", work_names)) #choosing only PC models because multicollinearity issue for boostrap

bootstrap_set <- model_train  |>
  bootstraps(times = 5)

workflows_bootstrap <- pc_workflows_tbl |>
  mutate(fits = list(fit_resamples(work_objects,
                                   bootstrap_set,
                                   metrics = consideration_metrics))) |>
  mutate(metrics = list(collect_metrics(fits)))
```

```
## > A | warning: prediction from a rank-deficient fit may be misleading
```

```
## There were issues with some computations   A: x1There were issues with some computations   A: x2There
## > A | warning: prediction from a rank-deficient fit may be misleading
## There were issues with some computations   A: x1There were issues with some computations   A: x2There
```

```
workflows_bootstrap
```

```
## # A tibble: 4 x 5
## # Rowwise:
```

```
##   work_names work_objects fits      predictions        metrics
##   <chr>      <list>       <list>    <list>             <list>
## 1 lm_4PC     <workflow>   <rsmp[+]> <tibble [516 x 1]> <tibble [3 x 6]>
## 2 lm_6PC     <workflow>   <rsmp[+]> <tibble [516 x 1]> <tibble [3 x 6]>
## 3 knn_4PC    <workflow>   <rsmp[+]> <tibble [516 x 1]> <tibble [3 x 6]>
## 4 rf_4PC     <workflow>   <rsmp[+]> <tibble [516 x 1]> <tibble [3 x 6]>
```

```
workflows_boot_results <- workflows_bootstrap |>
  select(work_names,
         metrics) |>
  unnest(metrics) |>
  select(work_names,
         mean) |>
  arrange(work_names)

workflows_boot_results
```

```
## # A tibble: 12 x 2
##    work_names          mean
##    <chr>              <dbl>
##  1 knn_4PC     90819773.
##  2 knn_4PC    629052828.
##  3 knn_4PC            0.877
##  4 lm_4PC      33533142.
##  5 lm_4PC      96964481.
##  6 lm_4PC             0.997
##  7 lm_6PC      27474007.
##  8 lm_6PC     117764367.
##  9 lm_6PC             0.995
## 10 rf_4PC      79441834.
## 11 rf_4PC     640365396.
## 12 rf_4PC            0.881
```

uncertainty quantification: held out test set

```
set.seed(1)
val_set <- validation_split(model_train,
                            prop = 0.75,
                            strata = Total.Consideration.for.Taxable.Conveyances)
```

```
## Warning: `validation_split()` was deprecated in rsample 1.2.0.
## i Please use `initial_validation_split()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
val_set |>
  class()
```

```
## [1] "validation_split" "rset"             "tbl_df"           "tbl"
## [5] "data.frame"
```

```
workflows_val <- workflows_tbl |>
  mutate(fits = list(fit_resamples(work_objects,
                                   val_set,
```

```
                                    metrics = consideration_metrics))) |>
  mutate(metrics = list(collect_metrics(fits)))
```

```
## > A | warning: prediction from a rank-deficient fit may be misleading
```

```
## There were issues with some computations   A: x1There were issues with some computations   A: x1
## > A | warning: prediction from a rank-deficient fit may be misleading
## There were issues with some computations   A: x1There were issues with some computations   A: x1
## > A | warning: prediction from a rank-deficient fit may be misleading
## There were issues with some computations   A: x1There were issues with some computations   A: x1
```

```
workflows_val
```

```
## # A tibble: 7 x 5
## # Rowwise:
##   work_names work_objects fits      predictions        metrics
##   <chr>      <list>       <list>    <list>             <list>
## 1 lm_4PC     <workflow>   <rsmp[+]> <tibble [516 x 1]> <tibble [3 x 6]>
## 2 lm_6PC     <workflow>   <rsmp[+]> <tibble [516 x 1]> <tibble [3 x 6]>
## 3 lm         <workflow>   <rsmp[+]> <tibble [516 x 1]> <tibble [3 x 6]>
## 4 knn_4PC    <workflow>   <rsmp[+]> <tibble [516 x 1]> <tibble [3 x 6]>
## 5 knn        <workflow>   <rsmp[+]> <tibble [516 x 1]> <tibble [3 x 6]>
## 6 rf_4PC     <workflow>   <rsmp[+]> <tibble [516 x 1]> <tibble [3 x 6]>
## 7 rf         <workflow>   <rsmp[+]> <tibble [516 x 1]> <tibble [3 x 6]>
```

```
workflows_val |>
  select(c(work_names,
           metrics)) |>
  unnest(metrics) |>
  arrange(.metric)
```

```
## # A tibble: 21 x 7
##    work_names .metric .estimator         mean     n std_err .config
##    <chr>      <chr>   <chr>             <dbl> <int>   <dbl> <chr>
##  1 lm_4PC     mae     standard   26588775.        1      NA Preprocessor1_Model1
##  2 lm_6PC     mae     standard   24203310.        1      NA Preprocessor1_Model1
##  3 lm         mae     standard          0.136     1      NA Preprocessor1_Model1
##  4 knn_4PC    mae     standard   67748143.        1      NA Preprocessor1_Model1
##  5 knn        mae     standard   65465395.        1      NA Preprocessor1_Model1
##  6 rf_4PC     mae     standard   58755568.        1      NA Preprocessor1_Model1
##  7 rf         mae     standard   33869773.        1      NA Preprocessor1_Model1
##  8 lm_4PC     rmse    standard   90240003.        1      NA Preprocessor1_Model1
##  9 lm_6PC     rmse    standard   86341036.        1      NA Preprocessor1_Model1
## 10 lm         rmse    standard          0.297     1      NA Preprocessor1_Model1
## # i 11 more rows
```

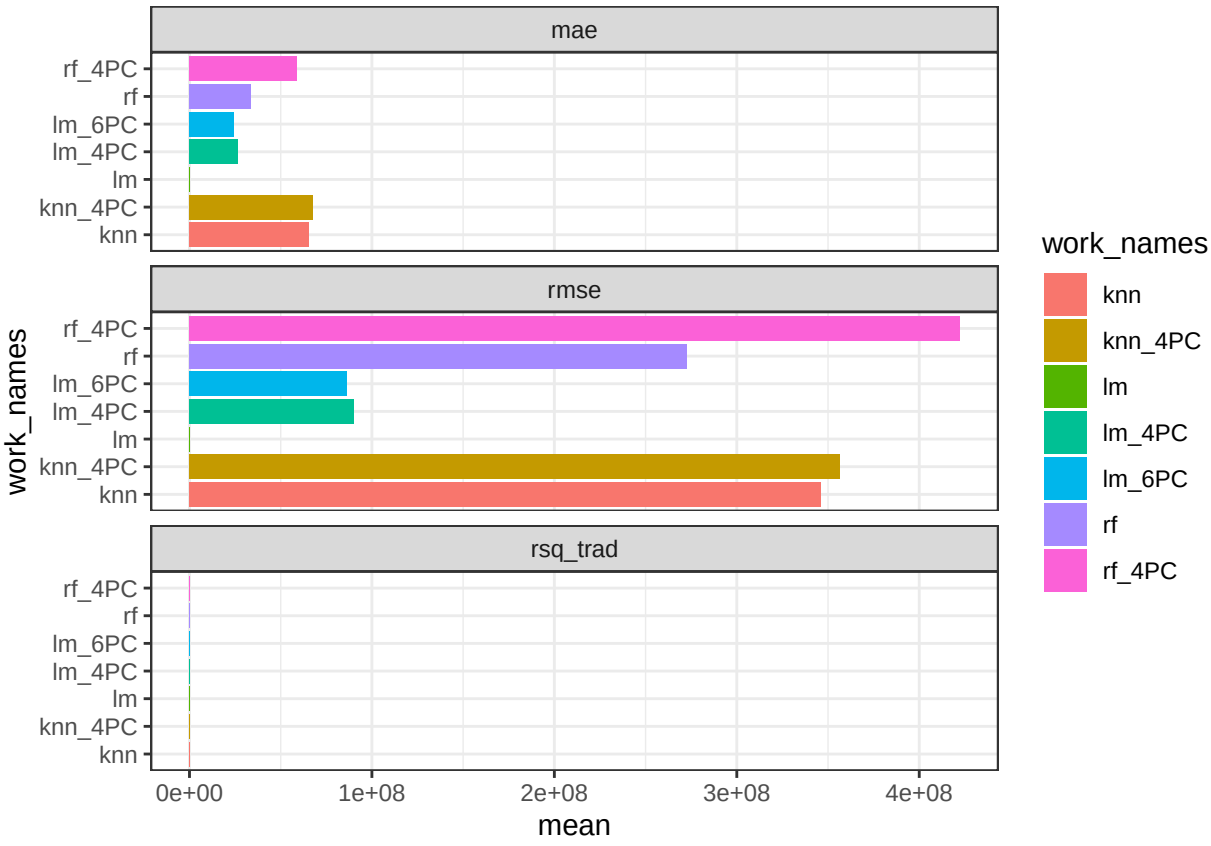
```
workflows_val |>
  select(c(work_names,
           metrics)) |>
  unnest(metrics)
```

```
## # A tibble: 21 x 7
##    work_names .metric  .estimator         mean     n std_err .config
##    <chr>      <chr>    <chr>             <dbl> <int>   <dbl> <chr>
##  1 lm_4PC     mae      standard   26588775.        1      NA Preprocessor1_Mode~
##  2 lm_4PC     rmse     standard   90240003.        1      NA Preprocessor1_Mode~
##  3 lm_4PC     rsq_trad standard          0.994     1      NA Preprocessor1_Mode~
```

```
##  4 lm_6PC     mae      standard   24203310.        1      NA Preprocessor1_Mode~
##  5 lm_6PC     rmse     standard   86341036.        1      NA Preprocessor1_Mode~
##  6 lm_6PC     rsq_trad standard          0.994     1      NA Preprocessor1_Mode~
##  7 lm         mae      standard          0.136     1      NA Preprocessor1_Mode~
##  8 lm         rmse     standard          0.297     1      NA Preprocessor1_Mode~
##  9 lm         rsq_trad standard          1         1      NA Preprocessor1_Mode~
## 10 knn_4PC    mae      standard   67748143.        1      NA Preprocessor1_Mode~
## # i 11 more rows
```

rsq_trad not missing, just super small

```
workflows_val |>
  select(c(work_names,
           metrics)) |>
  unnest(metrics) |>
  ggplot(aes(y = work_names,
             fill = work_names,
             x = mean)) +
  geom_col() +
  facet_wrap(~.metric,
             nrow = 3) #rsq_trad not missing, just super small
```



Add ons we used:

- Perform Principal Component Analysis
- Impute missing data
- Make use of a list column during your analysis

- Make a github repository for your project (share the link with me)
- Include more than one form of uncertainty quantification from the bulleted list.