01_clean_data.Rmd

Vinay K L

2023-10-07

Loading the data

```
data1 <- read.csv("../data/vertnet_latest_birds.csv", sep = ",", header = TRUE)
```

```
data2 <- read.csv("../data/aves_vertnet_records-43452c1446904be6ba9641c86f891234.tsv",
    sep = "\t", header = TRUE)
```

inspecting the dataframe

```
head(data1)
```

```
##   beginrecord icode                                  title
## 1       begin    AM Australian Museum provider for OZCAM
## 2       begin    AM Australian Museum provider for OZCAM
## 3       begin    AM Australian Museum provider for OZCAM
## 4       begin    AM Australian Museum provider for OZCAM
## 5       begin    AM Australian Museum provider for OZCAM
## 6       begin    AM Australian Museum provider for OZCAM
##
## 1 Australian Museum. Australian Museum provider for OZCAM. Source: http://collections.ala.org.au/publ
## 2 Australian Museum. Australian Museum provider for OZCAM. Source: http://collections.ala.org.au/publ
## 3 Australian Museum. Australian Museum provider for OZCAM. Source: http://collections.ala.org.au/publ
## 4 Australian Museum. Australian Museum provider for OZCAM. Source: http://collections.ala.org.au/publ
## 5 Australian Museum. Australian Museum provider for OZCAM. Source: http://collections.ala.org.au/publ
## 6 Australian Museum. Australian Museum provider for OZCAM. Source: http://collections.ala.org.au/publ
##           contact                email
## 1 OZCAM Webmaster OZCAM.CHAFC@gmail.com
## 2 OZCAM Webmaster OZCAM.CHAFC@gmail.com
## 3 OZCAM Webmaster OZCAM.CHAFC@gmail.com
## 4 OZCAM Webmaster OZCAM.CHAFC@gmail.com
## 5 OZCAM Webmaster OZCAM.CHAFC@gmail.com
## 6 OZCAM Webmaster OZCAM.CHAFC@gmail.com
##                                      emlrights
## 1 http://creativecommons.org/licenses/by/3.0/au/
## 2 http://creativecommons.org/licenses/by/3.0/au/
## 3 http://creativecommons.org/licenses/by/3.0/au/
## 4 http://creativecommons.org/licenses/by/3.0/au/
## 5 http://creativecommons.org/licenses/by/3.0/au/
```

```
## 6 http://creativecommons.org/licenses/by/3.0/au/
##                          gbifdatasetid                       gbifpublisherid doi
## 1 dce8feb0-6c89-11de-8225-b8a03c50a862 770c30d2-c2a8-4bb2-8056-6167297cddae
## 2 dce8feb0-6c89-11de-8225-b8a03c50a862 770c30d2-c2a8-4bb2-8056-6167297cddae
## 3 dce8feb0-6c89-11de-8225-b8a03c50a862 770c30d2-c2a8-4bb2-8056-6167297cddae
## 4 dce8feb0-6c89-11de-8225-b8a03c50a862 770c30d2-c2a8-4bb2-8056-6167297cddae
## 5 dce8feb0-6c89-11de-8225-b8a03c50a862 770c30d2-c2a8-4bb2-8056-6167297cddae
## 6 dce8feb0-6c89-11de-8225-b8a03c50a862 770c30d2-c2a8-4bb2-8056-6167297cddae
##     migrator                          networks orgcountry           orgname
## 1 2015-01-05 MaNIS,ORNIS,HerpNET,VertNet,OZCAM  Australia Australian Museum
## 2 2015-01-05 MaNIS,ORNIS,HerpNET,VertNet,OZCAM  Australia Australian Museum
## 3 2015-01-05 MaNIS,ORNIS,HerpNET,VertNet,OZCAM  Australia Australian Museum
## 4 2015-01-05 MaNIS,ORNIS,HerpNET,VertNet,OZCAM  Australia Australian Museum
## 5 2015-01-05 MaNIS,ORNIS,HerpNET,VertNet,OZCAM  Australia Australian Museum
## 6 2015-01-05 MaNIS,ORNIS,HerpNET,VertNet,OZCAM  Australia Australian Museum
##   orgstateprovince    pubdate                                       source_url
## 1  New South Wales 2015-01-07 http://collections.ala.org.au/public/show/dr340
## 2  New South Wales 2015-01-07 http://collections.ala.org.au/public/show/dr340
## 3  New South Wales 2015-01-07 http://collections.ala.org.au/public/show/dr340
## 4  New South Wales 2015-01-07 http://collections.ala.org.au/public/show/dr340
## 5  New South Wales 2015-01-07 http://collections.ala.org.au/public/show/dr340
## 6  New South Wales 2015-01-07 http://collections.ala.org.au/public/show/dr340
##                          iptrecordid associatedmedia associatedoccurrences
## 1 5bf8e6e4-3db4-4ed6-a233-42a59729470a
## 2 f4f473f1-9064-4511-8c03-2ea0f676e63c
## 3 19650040-9c7c-4110-8d8f-107f27faacea
## 4 6d297bac-ca80-45b0-982e-8bea1003bd3d
## 5 2452e6b6-db78-4320-8fc8-85299b1107ec
## 6 b7fb8471-1980-4081-9a99-b31f39dfd106
##   associatedorganisms associatedreferences associatedsequences associatedtaxa
## 1                  NA
## 2                  NA
## 3                  NA
## 4                  NA
## 5                  NA
## 6                  NA
##   bed behavior catalognumber   continent coordinateprecision
## 1                    O.74675
## 2                    O.74105 Australasia               0.001
## 3                    O.65233                           0.001
## 4                      A.669     Oceania               0.001
## 5                     A.6489 Australasia               0.001
## 6                    O.11509 Australasia               0.001
##   coordinateuncertaintyinmeters   country countrycode county dateidentified day
## 1                            NA                                             NA
## 2                            NA Australia          AU                       NA
## 3                         10000                                             NA
## 4                        100000   Vanuatu          VU                       NA
## 5                         10000 Australia          AU                       NA
## 6                         10000 Australia          AU                       NA
##   decimallatitude decimallongitude disposition earliestageorloweststage
## 1              NA               NA
## 2         -31.550          159.083
## 3           7.000         -147.500
```

```
## 4         -18.500          169.333
## 5         -41.100          146.816
## 6         -36.616          143.266
##   earliesteonorlowesteonothem earliestepochorlowestseries
## 1
## 2
## 3
## 4
## 5
## 6
##   earliesteraorlowesterathem earliestperiodorlowestsystem enddayofyear
## 1                                                                   NA
## 2                                                                   NA
## 3                                                                   NA
## 4                                                                   NA
## 5                                                                   NA
## 6                                                                   NA
##   establishmentmeans eventdate                                 eventid
## 1                            urn:australianmuseum.net.au:Events:3029244
## 2                            urn:australianmuseum.net.au:Events:1169708
## 3                            urn:australianmuseum.net.au:Events:3016458
## 4                            urn:australianmuseum.net.au:Events:3017197
## 5                            urn:australianmuseum.net.au:Events:3022467
## 6                            urn:australianmuseum.net.au:Events:3003881
##   eventremarks eventtime fieldnotes fieldnumber footprintspatialfit
## 1                                                               NA
## 2                                                               NA
## 3                                                               NA
## 4                                                               NA
## 5                                                               NA
## 6                                                               NA
##   footprintsrs footprintwkt formation               geodeticdatum
## 1           NA           NA
## 2           NA           NA           not recorded (forced WGS84)
## 3           NA           NA           not recorded (forced WGS84)
## 4           NA           NA           not recorded (forced WGS84)
## 5           NA           NA           not recorded (forced WGS84)
## 6           NA           NA           not recorded (forced WGS84)
##   geologicalcontextid georeferencedby georeferenceddate georeferenceprotocol
## 1
## 2
## 3
## 4
## 5
## 6
##   georeferenceremarks georeferencesources georeferenceverificationstatus group
## 1
## 2                                                 requires verification
## 3                                                 requires verification
## 4                                                 requires verification
## 5                                                 requires verification
## 6                                                 requires verification
##   habitat                            highergeography highergeographyid
## 1                                                                   NA
```

```
## 2         | Australia | New South Wales |  |  |  |  |                NA
## 3                                                                       NA
## 4                           | Vanuatu |  |  |  |  |  |                NA
## 5                | Australia | Tasmania |  |  |  |  |                NA
## 6                | Australia | Victoria |  |  |  |  |                NA
##   highestbiostratigraphiczone identificationid identificationqualifier
## 1                                          NA
## 2                                          NA
## 3                                          NA
## 4                                          NA
## 5                                          NA
## 6                                          NA
##   identificationreferences identificationremarks
## 1
## 2
## 3
## 4
## 5
## 6
##   identificationverificationstatus identifiedby individualcount island
## 1
## 2
## 3
## 4
## 5
## 6
##   islandgroup latestageorhighereststage latesteonorhighesteonothem
## 1
## 2
## 3
## 4     Vanuatu
## 5
## 6
##   latestepochorhighestseries latesteraorhighesterathem
## 1
## 2
## 3
## 4
## 5
## 6
##   latestperiodorhighestsystem lifestage lithostratigraphicterms locality
## 1
## 2
## 3
## 4
## 5
## 6
##   locationaccordingto locationid
## 1
## 2
## 3
## 4
## 5
## 6
```

```
##
## 1                                                                    "ecatalogue.LocCollectionEventLocal
## 2                                                                "ecatalogue.LocCollectionEventLocal: "Austra
## 3                                                    "ecatalogue.LocCollectionEventLocal: "Pacific Ocean, STATION #10 (7° 0
## 4                                          "ecatalogue.LocCollectionEventLocal: "Vanuatu, TANNA IS (18° 30' S, 169° 20' E
## 5 "ecatalogue.LocCollectionEventLocal: "Australia, Tasmania, GEORGE TOWN AREA (41° 06' S, 146° 49' E
## 6                                        "ecatalogue.LocCollectionEventLocal: "Australia, Victoria, ST ARN
##   lowestbiostratigraphiczone materialsampleid maximumdepthinmeters
## 1                                         NA                   NA
## 2                                         NA                   NA
## 3                                         NA                   NA
## 4                                         NA                   NA
## 5                                         NA                   NA
## 6                                         NA                   NA
##   maximumdistanceabovesurfaceinmeters maximumelevationinmeters member
## 1                                  NA                       NA
## 2                                  NA                       NA
## 3                                  NA                       NA
## 4                                  NA                       NA
## 5                                  NA                       NA
## 6                                  NA                       NA
##   minimumdepthinmeters minimumdistanceabovesurfaceinmeters
## 1                   NA                                 NA
## 2                   NA                                 NA
## 3                   NA                                 NA
## 4                   NA                                 NA
## 5                   NA                                 NA
## 6                   NA                                 NA
##   minimumelevationinmeters month municipality
## 1
## 2
## 3
## 4
## 5
## 6
##                          occurrenceid occurrenceremarks occurrencestatus
## 1 5bf8e6e4-3db4-4ed6-a233-42a59729470a                           present
## 2 f4f473f1-9064-4511-8c03-2ea0f676e63c                           present
## 3 19650040-9c7c-4110-8d8f-107f27faacea                           present
## 4 6d297bac-ca80-45b0-982e-8bea1003bd3d                           present
## 5 2452e6b6-db78-4320-8fc8-85299b1107ec                           present
## 6 b7fb8471-1980-4081-9a99-b31f39dfd106                           present
##   organismid organismname organismremarks organismscope othercatalognumbers
## 1                     NA              NA             NA
## 2                     NA              NA             NA
## 3                     NA              NA             NA
## 4                     NA              NA             NA
## 5                     NA              NA             NA
## 6                     NA              NA             NA
##   pointradiusspatialfit preparations previousidentifications
## 1                    NA
## 2                    NA
## 3                    NA
## 4                    NA
```

```
## 5                    NA
## 6                    NA
##                            recordedby recordnumber reproductivecondition
## 1                       HALLSTROM, E. J.
## 2
## 3                 SMITHSONIAN INSTITUTION
## 4 MACKINLAY, DR. ARCHIBALD - HMS NYMPHE
## 5                    BROADBENT, KENDALL
## 6                           GABRIEL, J.
##   samplingeffort samplingprotocol sex startdayofyear   stateprovince typestatus
## 1             NA                                 NA
## 2             NA                                 NA New South Wales
## 3             NA                                 NA
## 4             NA                                 NA
## 5             NA                                 NA        Tasmania
## 6             NA                                 NA        Victoria
##   verbatimcoordinates verbatimcoordinatesystem verbatimdepth verbatimelevation
## 1
## 2
## 3
## 4
## 5
## 6
##   verbatimeventdate verbatimlatitude            verbatimlocality
## 1
## 2                                   Australia | New South Wales
## 3
## 4                                                       Vanuatu
## 5                                         Australia | Tasmania
## 6                                         Australia | Victoria
##   verbatimlongitude verbatimsrs           waterbody year         dctype
## 1                             NA                        NA PhysicalObject
## 2                             NA                        NA PhysicalObject
## 3                             NA                        NA PhysicalObject
## 4                             NA South Pacific Ocean    NA PhysicalObject
## 5                             NA                        NA PhysicalObject
## 6                             NA                        NA PhysicalObject
##     modified language license rightsholder
## 1 2015-01-07       en    CCBY
## 2 2015-01-07       en    CCBY
## 3 2015-01-07       en    CCBY
## 4 2015-01-07       en    CCBY
## 5 2015-01-07       en    CCBY
## 6 2015-01-07       en    CCBY
##                              accessrights
## 1 http://vertnet.org/resources/norms.html
## 2 http://vertnet.org/resources/norms.html
## 3 http://vertnet.org/resources/norms.html
## 4 http://vertnet.org/resources/norms.html
## 5 http://vertnet.org/resources/norms.html
## 6 http://vertnet.org/resources/norms.html
##
## 1 Australian Museum. Australian Museum provider for OZCAM. Record ID: 5bf8e6e4-3db4-4ed6-a233-42a597
## 2 Australian Museum. Australian Museum provider for OZCAM. Record ID: f4f473f1-9064-4511-8c03-2ea0f6
```

```
## 3 Australian Museum. Australian Museum provider for OZCAM. Record ID: 19650040-9c7c-4110-8d8f-107f27
## 4 Australian Museum. Australian Museum provider for OZCAM. Record ID: 6d297bac-ca80-45b0-982e-8bea10
## 5 Australian Museum. Australian Museum provider for OZCAM. Record ID: 2452e6b6-db78-4320-8fc8-85299b
## 6 Australian Museum. Australian Museum provider for OZCAM. Record ID: b7fb8471-1980-4081-9a99-b31f39
##                                                                    dc_references
## 1 http://portal.vertnet.org/o/am/ornithology?id=5bf8e6e4-3db4-4ed6-a233-42a59729470a
## 2 http://portal.vertnet.org/o/am/ornithology?id=f4f473f1-9064-4511-8c03-2ea0f676e63c
## 3 http://portal.vertnet.org/o/am/ornithology?id=19650040-9c7c-4110-8d8f-107f27faacea
## 4 http://portal.vertnet.org/o/am/ornithology?id=6d297bac-ca80-45b0-982e-8bea1003bd3d
## 5 http://portal.vertnet.org/o/am/ornithology?id=2452e6b6-db78-4320-8fc8-85299b1107ec
## 6 http://portal.vertnet.org/o/am/ornithology?id=b7fb8471-1980-4081-9a99-b31f39dfd106
##   institutionid collectionid datasetid institutioncode collectioncode
## 1                                                  AM    Ornithology
## 2                                                  AM    Ornithology
## 3                                                  AM    Ornithology
## 4                                                  AM    Ornithology
## 5                                                  AM    Ornithology
## 6                                                  AM    Ornithology
##   datasetname ownerinstitutioncode     basisofrecord informationwithheld
## 1                                PreservedSpecimen
## 2                                PreservedSpecimen
## 3                                PreservedSpecimen
## 4                                PreservedSpecimen
## 5                                PreservedSpecimen
## 6                                PreservedSpecimen
##   datageneralizations dynamicproperties scientificnameid namepublishedinid
## 1
## 2
## 3
## 4
## 5
## 6
##               scientificname acceptednameusage originalnameusage
## 1              Ara militaris                NA
## 2             Anous stolidus                NA
## 3      Oceanodroma leucorhoa                NA
## 4        Myzomela cardinalis                NA
## 5 Melithreptus validirostris                NA
## 6           Myiagra inquieta                NA
##   namepublishedin namepublishedinyear
## 1                                  NA
## 2                                  NA
## 3                                  NA
## 4                                  NA
## 5                                  NA
## 6                                  NA
##                                                 higherclassification
## 1             Animalia | Chordata |  | Psittaciformes | Psittacidae | Ara
## 2              Animalia | Chordata |  | Charadriiformes | Laridae | Anous
## 3 Animalia | Chordata |  | Procellariiformes | Hydrobatidae | Oceanodroma
## 4        Animalia | Chordata |  | Passeriformes | Meliphagidae | Myzomela
## 5    Animalia | Chordata |  | Passeriformes | Meliphagidae | Melithreptus
## 6          Animalia | Chordata |  | Passeriformes | Monarchidae | Myiagra
##    kingdom   phylum class             order       family        genus subgenus
```

```
## 1 Animalia Chordata  Aves    Psittaciformes  Psittacidae          Ara
## 2 Animalia Chordata  Aves   Charadriiformes      Laridae        Anous
## 3 Animalia Chordata  Aves Procellariiformes Hydrobatidae  Oceanodroma
## 4 Animalia Chordata  Aves     Passeriformes Meliphagidae     Myzomela
## 5 Animalia Chordata  Aves     Passeriformes Meliphagidae Melithreptus
## 6 Animalia Chordata  Aves     Passeriformes  Monarchidae      Myiagra
##   specificepithet infraspecificepithet taxonrank verbatimtaxonrank
## 1       militaris                       species
## 2        stolidus                       species
## 3       leucorhoa                       species
## 4      cardinalis                       species
## 5   validirostris                       species
## 6        inquieta                       species
##   scientificnameauthorship vernacularname nomenclaturalcode taxonomicstatus
## 1                                                     ICZN
## 2                          Common Noddy               ICZN
## 3                                                     ICZN
## 4                                                     ICZN
## 5                                                     ICZN
## 6                                                     ICZN
##                                                 keyname haslicense   vntype rank
## 1 am/ornithology/5bf8e6e4-3db4-4ed6-a233-42a59729470a          1 specimen    1
## 2 am/ornithology/f4f473f1-9064-4511-8c03-2ea0f676e63c          1 specimen    5
## 3 am/ornithology/19650040-9c7c-4110-8d8f-107f27faacea          1 specimen    9
## 4 am/ornithology/6d297bac-ca80-45b0-982e-8bea1003bd3d          1 specimen    9
## 5 am/ornithology/2452e6b6-db78-4320-8fc8-85299b1107ec          1 specimen    9
## 6 am/ornithology/b7fb8471-1980-4081-9a99-b31f39dfd106          1 specimen    9
##   mappable hashid hastypestatus wascaptive wasinvasive hastissue hasmedia
## 1        0   8301             0          0           0         0        0
## 2        1   5435             0          0           0         0        0
## 3        1   9408             0          0           0         0        0
## 4        1   2766             0          0           0         0        0
## 5        1   1112             0          0           0         0        0
## 6        1   8555             0          0           0         0        0
##   isfossil haslength haslifestage hasmass hassex lengthinmm massing
## 1        0         0            0       0      0         NA      NA
## 2        0         0            0       0      0         NA      NA
## 3        0         0            0       0      0         NA      NA
## 4        0         0            0       0      0         NA      NA
## 5        0         0            0       0      0         NA      NA
## 6        0         0            0       0      0         NA      NA
##   lengthunitsinferred massunitsinferred underivedlifestage underivedsex
## 1                  NA                NA
## 2                  NA                NA
## 3                  NA                NA
## 4                  NA                NA
## 5                  NA                NA
## 6                  NA                NA
```

```
summary(data1)
```

```
##  beginrecord           icode              title              citation
##  Length:1386455     Length:1386455     Length:1386455     Length:1386455
##  Class :character   Class :character   Class :character   Class :character
```

```
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    contact             email            emlrights         gbifdatasetid
##  Length:1386455     Length:1386455     Length:1386455     Length:1386455
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  gbifpublisherid        doi              migrator           networks
##  Length:1386455     Length:1386455     Length:1386455     Length:1386455
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   orgcountry          orgname          orgstateprovince     pubdate
##  Length:1386455     Length:1386455     Length:1386455     Length:1386455
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   source_url         iptrecordid        associatedmedia    associatedoccurrences
##  Length:1386455     Length:1386455     Length:1386455     Length:1386455
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  associatedorganisms associatedreferences associatedsequences
##  Mode:logical        Length:1386455       Length:1386455
##  NA's:1386455        Class :character     Class :character
##                      Mode  :character     Mode  :character
##
##
##
##
##  associatedtaxa         bed              behavior         catalognumber
##  Length:1386455     Length:1386455     Length:1386455     Length:1386455
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   continent         coordinateprecision coordinateuncertaintyinmeters
```

```
##  Length:1386455     Length:1386455      Min.   :       1
##  Class :character   Class :character    1st Qu.:    1610
##  Mode  :character   Mode  :character    Median :    3370
##                                         Mean   :   28850
##                                         3rd Qu.:   10000
##                                         Max.   :96597547
##                                         NA's   :1142385
##    country          countrycode           county          dateidentified
##  Length:1386455     Length:1386455     Length:1386455     Length:1386455
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##       day          decimallatitude  decimallongitude disposition
##  Min.   :   0.00   Min.   :-78.3    Min.   :-180.0   Length:1386455
##  1st Qu.:   8.00   1st Qu.:  3.0    1st Qu.:-110.7   Class :character
##  Median :  15.00   Median : 31.9    Median : -83.0   Mode  :character
##  Mean   :  15.81   Mean   : 21.6    Mean   : -43.3
##  3rd Qu.:  23.00   3rd Qu.: 41.3    3rd Qu.:  11.2
##  Max.   :1999.00   Max.   : 86.2    Max.   : 180.0
##  NA's   :252549    NA's   :891549   NA's   :891549
##  earliestageorloweststage earliesteonorlowesteonothem
##  Length:1386455           Length:1386455
##  Class :character         Class :character
##  Mode  :character         Mode  :character
##
##
##
##
##  earliestepochorlowestseries earliesteraorlowesterathem
##  Length:1386455              Length:1386455
##  Class :character            Class :character
##  Mode  :character            Mode  :character
##
##
##
##
##  earliestperiodorlowestsystem  enddayofyear    establishmentmeans
##  Length:1386455                Min.   :  1.0    Length:1386455
##  Class :character              1st Qu.:105.0    Class :character
##  Mode  :character              Median :164.0    Mode  :character
##                                Mean   :174.8
##                                3rd Qu.:251.0
##                                Max.   :366.0
##                                NA's   :296153
##   eventdate           eventid          eventremarks        eventtime
##  Length:1386455     Length:1386455     Length:1386455     Length:1386455
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
##
##   fieldnotes        fieldnumber        footprintspatialfit footprintsrs
##  Length:1386455     Length:1386455     Mode:logical        Mode:logical
##  Class :character   Class :character   NA's:1386455        NA's:1386455
##  Mode  :character   Mode  :character
##
##
##
##
##  footprintwkt    formation          geodeticdatum      geologicalcontextid
##  Mode:logical   Length:1386455     Length:1386455     Length:1386455
##  NA's:1386455   Class :character   Class :character   Class :character
##                 Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  georeferencedby    georeferenceddate  georeferenceprotocol georeferenceremarks
##  Length:1386455     Length:1386455     Length:1386455       Length:1386455
##  Class :character   Class :character   Class :character     Class :character
##  Mode  :character   Mode  :character   Mode  :character     Mode  :character
##
##
##
##
##  georeferencesources georeferenceverificationstatus    group
##  Length:1386455      Length:1386455                 Length:1386455
##  Class :character    Class :character               Class :character
##  Mode  :character    Mode  :character               Mode  :character
##
##
##
##
##    habitat          highergeography    highergeographyid
##  Length:1386455     Length:1386455     Mode:logical
##  Class :character   Class :character   NA's:1386455
##  Mode  :character   Mode  :character
##
##
##
##
##  highestbiostratigraphiczone identificationid identificationqualifier
##  Length:1386455              Mode:logical     Length:1386455
##  Class :character            NA's:1386455     Class :character
##  Mode  :character                             Mode  :character
##
##
##
##
##  identificationreferences identificationremarks
##  Length:1386455           Length:1386455
##  Class :character         Class :character
##  Mode  :character         Mode  :character
##
```

```
##
##
##
##  identificationverificationstatus identifiedby       individualcount
##  Length:1386455                   Length:1386455     Length:1386455
##  Class :character                 Class :character   Class :character
##  Mode  :character                 Mode  :character   Mode  :character
##
##
##
##
##     island          islandgroup        latestageorhighheststage
##  Length:1386455     Length:1386455     Length:1386455
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  latesteonorhighesteonothem latestepochorhighestseries
##  Length:1386455             Length:1386455
##  Class :character           Class :character
##  Mode  :character           Mode  :character
##
##
##
##
##  latesteraorhighesterathem latestperiodorhighestsystem  lifestage
##  Length:1386455            Length:1386455               Length:1386455
##  Class :character          Class :character             Class :character
##  Mode  :character          Mode  :character             Mode  :character
##
##
##
##
##  lithostratigraphicterms   locality         locationaccordingto
##  Length:1386455          Length:1386455     Length:1386455
##  Class :character        Class :character   Class :character
##  Mode  :character        Mode  :character   Mode  :character
##
##
##
##
##   locationid        locationremarks    lowestbiostratigraphiczone
##  Length:1386455     Length:1386455     Length:1386455
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  materialsampleid maximumdepthinmeters maximumdistanceabovesurfaceinmeters
##  Mode:logical     Min.   :1.0          Mode:logical
##  NA's:1386455     1st Qu.:1.8          NA's:1386455
```

```
##                   Median :4.5
##                   Mean   :4.8
##                   3rd Qu.:7.5
##                   Max.   :9.0
##                   NA's   :1386451
##  maximumelevationinmeters    member          minimumdepthinmeters
##  Min.   :     -76         Length:1386455     Min.   :1.0
##  1st Qu.:     420         Class :character   1st Qu.:1.8
##  Median :    1219         Mode  :character   Median :4.5
##  Mean   :    1631                            Mean   :4.8
##  3rd Qu.:    1890                            3rd Qu.:7.5
##  Max.   :14501300                            Max.   :9.0
##  NA's   :1306636                             NA's   :1386451
##  minimumdistanceabovesurfaceinmeters minimumelevationinmeters
##  Mode:logical                        Length:1386455
##  NA's:1386455                        Class :character
##                                      Mode  :character
##
##
##
##
##     month           municipality       occurrenceid       occurrenceremarks
##  Length:1386455     Length:1386455     Length:1386455     Length:1386455
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  occurrencestatus    organismid         organismname   organismremarks
##  Length:1386455     Length:1386455     Mode:logical   Mode:logical
##  Class :character   Class :character   NA's:1386455   NA's:1386455
##  Mode  :character   Mode  :character
##
##
##
##
##  organismscope  othercatalognumbers pointradiusspatialfit preparations
##  Mode:logical   Length:1386455      Mode:logical          Length:1386455
##  NA's:1386455   Class :character    NA's:1386455          Class :character
##                 Mode  :character                          Mode  :character
##
##
##
##
##  previousidentifications  recordedby         recordnumber
##  Length:1386455          Length:1386455     Length:1386455
##  Class :character        Class :character   Class :character
##  Mode  :character        Mode  :character   Mode  :character
##
##
##
##
##  reproductivecondition samplingeffort samplingprotocol       sex
```

```
##  Length:1386455        Mode:logical   Length:1386455     Length:1386455
##  Class :character      NA's:1386455   Class :character   Class :character
##  Mode  :character                     Mode  :character   Mode  :character
##
##
##
##
##  startdayofyear   stateprovince       typestatus        verbatimcoordinates
##  Min.   :  1.0    Length:1386455     Length:1386455     Length:1386455
##  1st Qu.:105.0    Class :character   Class :character   Class :character
##  Median :164.0    Mode  :character   Mode  :character   Mode  :character
##  Mean   :174.8
##  3rd Qu.:251.0
##  Max.   :366.0
##  NA's   :296153
##  verbatimcoordinatesystem verbatimdepth      verbatimelevation
##  Length:1386455           Length:1386455     Length:1386455
##  Class :character         Class :character   Class :character
##  Mode  :character         Mode  :character   Mode  :character
##
##
##
##
##  verbatimeventdate  verbatimlatitude   verbatimlocality   verbatimlongitude
##  Length:1386455     Length:1386455     Length:1386455     Length:1386455
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  verbatimsrs     waterbody              year             dctype
##  Mode:logical   Length:1386455     Min.   :    -186   Length:1386455
##  NA's:1386455   Class :character   1st Qu.:    1911   Class :character
##                 Mode  :character   Median :    1937   Mode  :character
##                                    Mean   :    2437
##                                    3rd Qu.:    1970
##                                    Max.   :20112012
##                                    NA's   :187241
##    modified           language           license          rightsholder
##  Length:1386455     Length:1386455     Length:1386455     Length:1386455
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  accessrights       bibliographiccitation dc_references      institutionid
##  Length:1386455     Length:1386455        Length:1386455     Length:1386455
##  Class :character   Class :character      Class :character   Class :character
##  Mode  :character   Mode  :character      Mode  :character   Mode  :character
##
##
##
```

```
##
##  collectionid        datasetid         institutioncode    collectioncode
##  Length:1386455     Length:1386455     Length:1386455     Length:1386455
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  datasetname        ownerinstitutioncode basisofrecord      informationwithheld
##  Length:1386455     Length:1386455       Length:1386455     Length:1386455
##  Class :character   Class :character     Class :character   Class :character
##  Mode  :character   Mode  :character     Mode  :character   Mode  :character
##
##
##
##
##  datageneralizations dynamicproperties  scientificnameid   namepublishedinid
##  Length:1386455      Length:1386455     Length:1386455     Length:1386455
##  Class :character    Class :character   Class :character   Class :character
##  Mode  :character    Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  scientificname     acceptednameusage originalnameusage  namepublishedin
##  Length:1386455     Mode:logical      Length:1386455     Length:1386455
##  Class :character   NA's:1386455      Class :character   Class :character
##  Mode  :character                     Mode  :character   Mode  :character
##
##
##
##
##  namepublishedinyear higherclassification   kingdom             phylum
##  Min.   :1829        Length:1386455       Length:1386455     Length:1386455
##  1st Qu.:1966        Class :character     Class :character   Class :character
##  Median :1966        Mode  :character     Mode  :character   Mode  :character
##  Mean   :1939
##  3rd Qu.:1966
##  Max.   :1966
##  NA's   :1386450
##     class              order              family             genus
##  Length:1386455     Length:1386455     Length:1386455     Length:1386455
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    subgenus          specificepithet    infraspecificepithet  taxonrank
##  Length:1386455     Length:1386455     Length:1386455       Length:1386455
##  Class :character   Class :character   Class :character     Class :character
##  Mode  :character   Mode  :character   Mode  :character     Mode  :character
##
```

```
##
##
##
##  verbatimtaxonrank  scientificnameauthorship vernacularname
##  Length:1386455     Length:1386455           Length:1386455
##  Class :character   Class :character         Class :character
##  Mode  :character   Mode  :character         Mode  :character
##
##
##
##
##  nomenclaturalcode  taxonomicstatus      keyname            haslicense
##  Length:1386455     Length:1386455     Length:1386455     Min.   :0.0000
##  Class :character   Class :character   Class :character   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Mode  :character   Median :1.0000
##                                                           Mean   :0.5972
##                                                           3rd Qu.:1.0000
##                                                           Max.   :1.0000
##                                                           NA's   :1
##     vntype               rank           mappable         hashid
##  Length:1386455     Min.   : 0.000   Min.   :0.000   Min.   :   0
##  Class :character   1st Qu.: 4.000   1st Qu.:0.000   1st Qu.:2500
##  Mode  :character   Median : 4.000   Median :0.000   Median :4981
##                     Mean   : 5.524   Mean   :0.357   Mean   :4985
##                     3rd Qu.: 8.000   3rd Qu.:1.000   3rd Qu.:7462
##                     Max.   :12.000   Max.   :1.000   Max.   :9998
##                     NA's   :1        NA's   :1       NA's   :1
##  hastypestatus        wascaptive        wasinvasive   hastissue
##  Min.   :0.000000   Min.   :0.000000   Min.   :0    Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0    1st Qu.:0.00000
##  Median :0.000000   Median :0.000000   Median :0    Median :0.00000
##  Mean   :0.009482   Mean   :0.007286   Mean   :0    Mean   :0.09143
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0    3rd Qu.:0.00000
##  Max.   :1.000000   Max.   :1.000000   Max.   :0    Max.   :1.00000
##  NA's   :1          NA's   :1          NA's   :1    NA's   :1
##     hasmedia          isfossil        haslength        haslifestage
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00000   Median :0.0000   Median :0.00000   Median :0.0000
##  Mean   :0.03528   Mean   :0.1495   Mean   :0.01169   Mean   :0.2841
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##  NA's   :1         NA's   :1        NA's   :1         NA's   :1
##     hasmass            hassex         lengthinmm          massing
##  Min.   :0.00000   Min.   :0.0000   Min.   :       1   Min.   :0.000e+00
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:      99   1st Qu.:1.400e+01
##  Median :0.00000   Median :1.0000   Median :     149   Median :2.800e+01
##  Mean   :0.05818   Mean   :0.6553   Mean   :   47328   Mean   :1.456e+05
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:     224   3rd Qu.:7.400e+01
##  Max.   :1.00000   Max.   :1.0000   Max.   :96101218   Max.   :6.834e+09
##  NA's   :1         NA's   :1        NA's   :1370290    NA's   :1305831
##  lengthunitsinferred massunitsinferred underivedlifestage underivedsex
##  Min.   :0.0         Min.   :0.0       Length:1386455     Length:1386455
##  1st Qu.:0.0         1st Qu.:0.0       Class :character   Class :character
```

```
##  Median :0.0         Median :0.0       Mode  :character   Mode  :character
##  Mean   :0.4         Mean   :0.2
##  3rd Qu.:1.0         3rd Qu.:0.0
##  Max.   :1.0         Max.   :1.0
##  NA's   :1370290     NA's   :1305831
```

```
str(data1)
```

```
## 'data.frame':    1386455 obs. of  194 variables:
##  $ beginrecord                          : chr  "begin" "begin" "begin" "begin" ...
##  $ icode                                : chr  "AM" "AM" "AM" "AM" ...
##  $ title                                : chr  "Australian Museum provider for OZCAM" "Australian Muse
##  $ citation                             : chr  "Australian Museum. Australian Museum provider for OZCA
##  $ contact                              : chr  "OZCAM Webmaster" "OZCAM Webmaster" "OZCAM Webmaster" "
##  $ email                                : chr  "OZCAM.CHAFC@gmail.com" "OZCAM.CHAFC@gmail.com" "OZCAM.
##  $ emlrights                            : chr  "http://creativecommons.org/licenses/by/3.0/au/" "http:
##  $ gbifdatasetid                        : chr  "dce8feb0-6c89-11de-8225-b8a03c50a862" "dce8feb0-6c89-1
##  $ gbifpublisherid                      : chr  "770c30d2-c2a8-4bb2-8056-6167297cddae" "770c30d2-c2a8-4
##  $ doi                                  : chr  "" "" "" "" ...
##  $ migrator                             : chr  "2015-01-05" "2015-01-05" "2015-01-05" "2015-01-05" ...
##  $ networks                             : chr  "MaNIS,ORNIS,HerpNET,VertNet,OZCAM" "MaNIS,ORNIS,HerpNE
##  $ orgcountry                           : chr  "Australia" "Australia" "Australia" "Australia" ...
##  $ orgname                              : chr  "Australian Museum" "Australian Museum" "Australian Mus
##  $ orgstateprovince                     : chr  "New South Wales" "New South Wales" "New South Wales" "
##  $ pubdate                              : chr  "2015-01-07" "2015-01-07" "2015-01-07" "2015-01-07" ...
##  $ source_url                           : chr  "http://collections.ala.org.au/public/show/dr340" "http
##  $ iptrecordid                          : chr  "5bf8e6e4-3db4-4ed6-a233-42a59729470a" "f4f473f1-9064-4
##  $ associatedmedia                      : chr  "" "" "" "" ...
##  $ associatedoccurrences                : chr  "" "" "" "" ...
##  $ associatedorganisms                  : logi  NA NA NA NA NA NA ...
##  $ associatedreferences                 : chr  "" "" "" "" ...
##  $ associatedsequences                  : chr  "" "" "" "" ...
##  $ associatedtaxa                       : chr  "" "" "" "" ...
##  $ bed                                  : chr  "" "" "" "" ...
##  $ behavior                             : chr  "" "" "" "" ...
##  $ catalognumber                        : chr  "O.74675" "O.74105" "O.65233" "A.669" ...
##  $ continent                            : chr  "" "Australasia" "" "Oceania" ...
##  $ coordinateprecision                  : chr  "" "0.001" "0.001" "0.001" ...
##  $ coordinateuncertaintyinmeters        : int  NA NA 10000 100000 10000 10000 10000 100000 10000 10000
##  $ country                              : chr  "" "Australia" "" "Vanuatu" ...
##  $ countrycode                          : chr  "" "AU" "" "VU" ...
##  $ county                               : chr  "" "" "" "" ...
##  $ dateidentified                       : chr  "" "" "" "" ...
##  $ day                                  : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ decimallatitude                      : num  NA -31.6 7 -18.5 -41.1 ...
##  $ decimallongitude                     : num  NA 159 -148 169 147 ...
##  $ disposition                          : chr  "" "" "" "" ...
##  $ earliestageorloweststage             : chr  "" "" "" "" ...
##  $ earliesteonorlowesteonothem          : chr  "" "" "" "" ...
##  $ earliestepochorlowestseries          : chr  "" "" "" "" ...
##  $ earliesteraorlowesterathem           : chr  "" "" "" "" ...
##  $ earliestperiodorlowestsystem         : chr  "" "" "" "" ...
##  $ enddayofyear                         : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ establishmentmeans                   : chr  "" "" "" "" ...
```

```
##  $ eventdate                          : chr  "" "" "" "" ...
##  $ eventid                            : chr  "urn:australianmuseum.net.au:Events:3029244" "urn:austra
##  $ eventremarks                       : chr  "" "" "" "" ...
##  $ eventtime                          : chr  "" "" "" "" ...
##  $ fieldnotes                         : chr  "" "" "" "" ...
##  $ fieldnumber                        : chr  "" "" "" "" ...
##  $ footprintspatialfit                : logi  NA NA NA NA NA NA ...
##  $ footprintsrs                       : logi  NA NA NA NA NA NA ...
##  $ footprintwkt                       : logi  NA NA NA NA NA NA ...
##  $ formation                          : chr  "" "" "" "" ...
##  $ geodeticdatum                      : chr  "" "not recorded (forced WGS84)" "not recorded (forced W
##  $ geologicalcontextid                : chr  "" "" "" "" ...
##  $ georeferencedby                    : chr  "" "" "" "" ...
##  $ georeferenceddate                  : chr  "" "" "" "" ...
##  $ georeferenceprotocol               : chr  "" "" "" "" ...
##  $ georeferenceremarks                : chr  "" "" "" "" ...
##  $ georeferencesources                : chr  "" "" "" "" ...
##  $ georeferenceverificationstatus     : chr  "" "requires verification" "requires verification" "requ
##  $ group                              : chr  "" "" "" "" ...
##  $ habitat                            : chr  "" "" "" "" ...
##  $ highergeography                    : chr  "" "| Australia | New South Wales |  |  |  |  |" "" "| W
##  $ highergeographyid                  : logi  NA NA NA NA NA NA ...
##  $ highestbiostratigraphiczone        : chr  "" "" "" "" ...
##  $ identificationid                   : logi  NA NA NA NA NA NA ...
##  $ identificationqualifier            : chr  "" "" "" "" ...
##  $ identificationreferences           : chr  "" "" "" "" ...
##  $ identificationremarks              : chr  "" "" "" "" ...
##  $ identificationverificationstatus   : chr  "" "" "" "" ...
##  $ identifiedby                       : chr  "" "" "" "" ...
##  $ individualcount                    : chr  "" "" "" "" ...
##  $ island                             : chr  "" "" "" "" ...
##  $ islandgroup                        : chr  "" "" "" "Vanuatu" ...
##  $ latestageorhighheststage           : chr  "" "" "" "" ...
##  $ latesteonorhighesteonothem         : chr  "" "" "" "" ...
##  $ latestepochorhighestseries         : chr  "" "" "" "" ...
##  $ latesteraorhighesterathem          : chr  "" "" "" "" ...
##  $ latestperiodorhighestsystem        : chr  "" "" "" "" ...
##  $ lifestage                          : chr  "" "" "" "" ...
##  $ lithostratigraphicterms            : chr  "" "" "" "" ...
##  $ locality                           : chr  "" "" "" "" ...
##  $ locationaccordingto                : chr  "" "" "" "" ...
##  $ locationid                         : chr  "" "" "" "" ...
##  $ locationremarks                    : chr  "\"ecatalogue.LocCollectionEventLocal: \"EX-CAGE (AUSTRA
##  $ lowestbiostratigraphiczone         : chr  "" "" "" "" ...
##  $ materialsampleid                   : logi  NA NA NA NA NA NA ...
##  $ maximumdepthinmeters               : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ maximumdistanceabovesurfaceinmeters: logi  NA NA NA NA NA NA ...
##  $ maximumelevationinmeters           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ member                             : chr  "" "" "" "" ...
##  $ minimumdepthinmeters               : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ minimumdistanceabovesurfaceinmeters: logi  NA NA NA NA NA NA ...
##  $ minimumelevationinmeters           : chr  "" "" "" "" ...
##  $ month                              : chr  "" "" "" "" ...
##  $ municipality                       : chr  "" "" "" "" ...
```

```
##   [list output truncated]
```

Checking for columns which are missing in data2

```
diff_cols_df1 <- setdiff(names(data1), names(data2))
cat("Columns in data1 not in data2: ", paste(diff_cols_df1, collapse = ", "), "\n")
```

```
## Columns in data1 not in data2:  beginrecord, icode, title, citation, contact, email, emlrights, doi,
```

Checking for columns which are missing in data1

```
diff_cols_df2 <- setdiff(names(data2), names(data1))
cat("Columns in data2 not in data1: ", paste(diff_cols_df2, collapse = ", "), "\n")
```

```
## Columns in data2 not in data1:  type, references, taxonremarks, lengthtype, dataset_url, dataset_cita
```

```
diff_cols_df2
```

```
##  [1] "type"                  "references"            "taxonremarks"
##  [4] "lengthtype"            "dataset_url"           "dataset_citation"
##  [7] "dataset_contact_email" "dataset_contact"       "dataset_pubdate"
## [10] "lastindexed"           "migrator_version"
```

Merge data

Now that we know both the data sets have some missing columns, lets remove them so that they can be merged

```
common_cols <- intersect(names(data1), names(data2))

data1 <- data1[, common_cols]
data2 <- data2[, common_cols]

merged_df <- rbind(data1, data2)
```

Retain Unique records

checking for any duplicate records and keeping unique records since we merged two different source data sets

```
unique_data <- unique(merged_df)
```

removing records which are all the fossilrecords

```r
unique_data <- subset(unique_data, basisofrecord != "FossilSpecimen")
```

Removing Unnecessary rows

Let's try and remove question marked rows from only certain columns - like country/family/order

```r
columns_to_check <- c("country", "order", "family")
```

Let's write a function to identify "?"

```r
has_question_mark_in_specific_columns <- function(row) {
    any(sapply(row[columns_to_check], function(col) grepl("\\?", col)))
}
```

Applying function

```r
question_mark_rows <- apply(unique_data, 1, has_question_mark_in_specific_columns)
```

```r
cleaned_df <- unique_data[!question_mark_rows, ]
```

Okay!! this removed few thousand records and we have about 1.24mil records.

Dealing with dates

```r
cleaned_df$year <- gsub(".*([0-9]{4}).*", "\\1", cleaned_df$year)
cleaned_df$year <- as.numeric(cleaned_df$year)
```

Binning years into chunks of 5 year interval

```r
bin_width <- 5
year_range <- seq(1600, 2023, bin_width)
```

```r
cleaned_df$YearBin <- cut(cleaned_df$year, breaks = year_range, labels = year_range[1:(length(year_rang
    1)])
```

Dealing with Country names

It seems like country column also lot of issues - let's try and clean that up.

```r
unique_countries <- unique(cleaned_df$country)
```

```r
standardized_countries <- countrycode(unique_countries, origin = "country.name",
    destination = "iso3c")
```

```
country_mapping <- data.frame(Original = unique_countries, Standardized = standardized_countries)
```

But I am not sure how to deal with this. Will keep it as it as for now

lets get total count and group the countries

```
country_counts <- cleaned_df %>%
    group_by(country) %>%
    summarise(count = n())
```

just testing merging country name after visually inspecting.

#{r } result_df <- count_data %>% mutate(Standardized = case_when(country %in% c("USA", "United States", "USa", "United States of America", "UNITED STATES") ~ "United States of America",   TRUE ~ as.character(country)   )) %>% group_by(Standardized) %>% summarize(total_count = sum(Count)) #

let's write this as a working data for the time being

```
write.csv(cleaned_df, "../data/cleaned_vertnet.csv", row.names = FALSE)
```