Assignment 12 B: MongoDB

Joshua Conte

July 30, 2017

NoSQL Data Storage & Retrieval using MongoDB

For this assignment, I need to learn how to:

- install the MongoDB server
- insert data into MongoDB
- fetch data from MongoDB

Tasks

Before I begin the tasks I need to configure R studio with the parameters below:

```r
# clears the console in RStudio
cat("\014")
```

```r
# clears environment
rm(list = ls())

# Set working directory
setwd("C:/R/DA5020/Week_12/Assignment_12_B")
```

Task 1

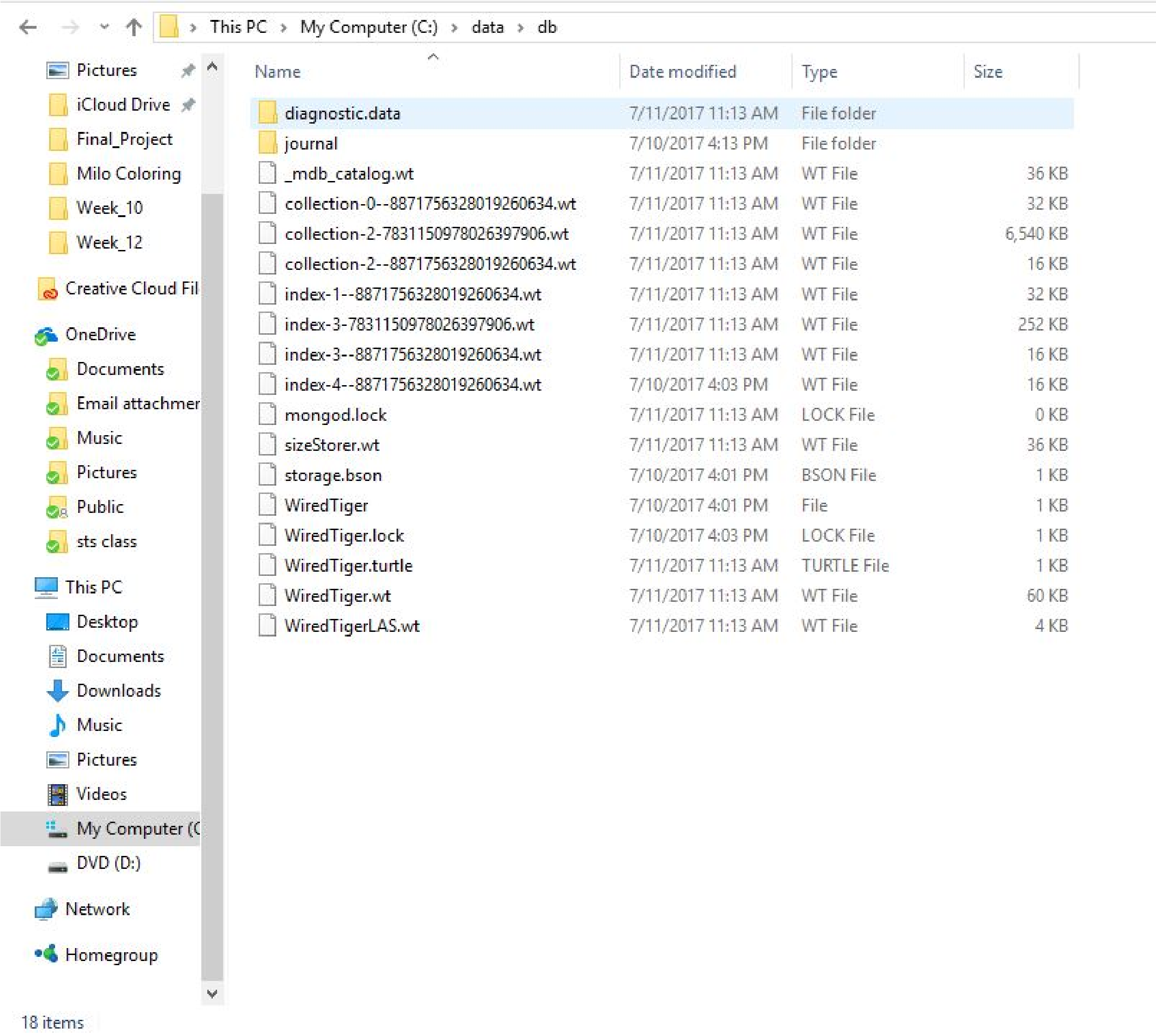
Install the MongoDB server on your system. Follow the step-by-step Guide to Install MongoDB from the official website as listed below or follow the instructions in Chapter 14 of the book.

a. Determine which MongoDB build you need.

I downloaded mongodb-win32-x86_64-2008plus-ssl-3.4.6-signed.msi and ran it as an executable.

b. Set up the MongoDB environment.

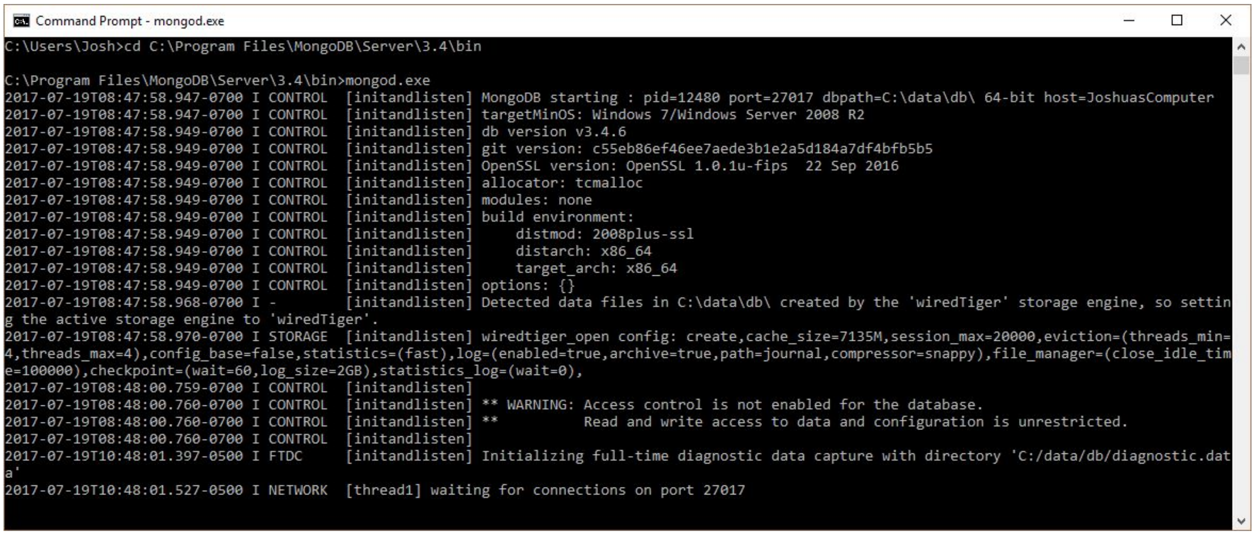
MongoDB requires a data directory to store all data. MongoDB's default data directory path is the absolute path C:/data/db. This can be made by opening the command prompt and typing "md C:/data/db". I created mine as shown in the screen shot below (I did not take a screenshot when I created it in the command prompt, but this is the result):



Name	Date modified	Type	Size
diagnostic.data	7/11/2017 11:13 AM	File folder	
journal	7/10/2017 4:13 PM	File folder	
_mdb_catalog.wt	7/11/2017 11:13 AM	WT File	36 KB
collection-0--8871756328019260634.wt	7/11/2017 11:13 AM	WT File	32 KB
collection-2-7831150978026397906.wt	7/11/2017 11:13 AM	WT File	6,540 KB
collection-2--8871756328019260634.wt	7/11/2017 11:13 AM	WT File	16 KB
index-1--8871756328019260634.wt	7/11/2017 11:13 AM	WT File	32 KB
index-3-7831150978026397906.wt	7/11/2017 11:13 AM	WT File	252 KB
index-3--8871756328019260634.wt	7/11/2017 11:13 AM	WT File	16 KB
index-4--8871756328019260634.wt	7/10/2017 4:03 PM	WT File	16 KB
mongod.lock	7/11/2017 11:13 AM	LOCK File	0 KB
sizeStorer.wt	7/11/2017 11:13 AM	WT File	36 KB
storage.bson	7/10/2017 4:01 PM	BSON File	1 KB
WiredTiger	7/10/2017 4:01 PM	File	1 KB
WiredTiger.lock	7/10/2017 4:03 PM	LOCK File	1 KB
WiredTiger.turtle	7/11/2017 11:13 AM	TURTLE File	1 KB
WiredTiger.wt	7/11/2017 11:13 AM	WT File	60 KB
WiredTigerLAS.wt	7/11/2017 11:13 AM	WT File	4 KB

c. Start MongoDB.

To start MongoDB, run mongod.exe, from the Command Prompt:



```
C:\Users\Josh>cd C:\Program Files\MongoDB\Server\3.4\bin

C:\Program Files\MongoDB\Server\3.4\bin>mongod.exe
2017-07-19T08:47:58.947-0700 I CONTROL  [initandlisten] MongoDB starting : pid=12480 port=27017 dbpath=C:\data\db\ 64-bit host=JoshuasComputer
2017-07-19T08:47:58.947-0700 I CONTROL  [initandlisten] targetMinOS: Windows 7/Windows Server 2008 R2
2017-07-19T08:47:58.949-0700 I CONTROL  [initandlisten] db version v3.4.6
2017-07-19T08:47:58.949-0700 I CONTROL  [initandlisten] git version: c55eb86ef46ee7aede3b1e2a5d184a7df4bfb5b5
2017-07-19T08:47:58.949-0700 I CONTROL  [initandlisten] OpenSSL version: OpenSSL 1.0.1u-fips  22 Sep 2016
2017-07-19T08:47:58.949-0700 I CONTROL  [initandlisten] allocator: tcmalloc
2017-07-19T08:47:58.949-0700 I CONTROL  [initandlisten] modules: none
2017-07-19T08:47:58.949-0700 I CONTROL  [initandlisten] build environment:
2017-07-19T08:47:58.949-0700 I CONTROL  [initandlisten]     distmod: 2008plus-ssl
2017-07-19T08:47:58.949-0700 I CONTROL  [initandlisten]     distarch: x86_64
2017-07-19T08:47:58.949-0700 I CONTROL  [initandlisten]     target_arch: x86_64
2017-07-19T08:47:58.949-0700 I CONTROL  [initandlisten] options: {}
2017-07-19T08:47:58.968-0700 I -        [initandlisten] Detected data files in C:\data\db\ created by the 'wiredTiger' storage engine, so settin
g the active storage engine to 'wiredTiger'.
2017-07-19T08:47:58.970-0700 I STORAGE  [initandlisten] wiredtiger_open config: create,cache_size=7135M,session_max=20000,eviction=(threads_min=
4,threads_max=4),config_base=false,statistics=(fast),log=(enabled=true,archive=true,path=journal,compressor=snappy),file_manager=(close_idle_tim
e=100000),checkpoint=(wait=60,log_size=2GB),statistics_log=(wait=0),
2017-07-19T08:48:00.759-0700 I CONTROL  [initandlisten]
2017-07-19T08:48:00.760-0700 I CONTROL  [initandlisten] ** WARNING: Access control is not enabled for the database.
2017-07-19T08:48:00.760-0700 I CONTROL  [initandlisten] **          Read and write access to data and configuration is unrestricted.
2017-07-19T08:48:00.760-0700 I CONTROL  [initandlisten]
2017-07-19T10:48:01.397-0500 I FTDC     [initandlisten] Initializing full-time diagnostic data capture with directory 'C:/data/db/diagnostic.dat
a'
2017-07-19T10:48:01.527-0500 I NETWORK  [thread1] waiting for connections on port 27017
```

This starts the main MongoDB database process. The waiting for connections message in the console output indicates that the mongod.exe process is running successfully.

Task 2

Insert the Bird Strikes.csv file into MongoDB and use the export command to display the inserted file. Note : Remember to reshape the data by removing the dots (periods) from the column names before inserting the data into MongoDB.

```r
# Load the data
# I also added NA to all blank cells to make it easier to analyze and
# stringsAsFactors = FALSE so I can remove levels from the data.
if (!exists("birdStrikes.df")) {
  birdStrikes.df <-
    read.csv(
      unz("Bird Strikes.zip", "Bird Strikes.csv"),
      header = TRUE,
      na.strings = c("", "NA"),
      stringsAsFactors = FALSE,
      sep = ","
    )
}

# Remove dots from names
names(birdStrikes.df) <- gsub("\\.", "", names(birdStrikes.df))
names(birdStrikes.df)
```

```
##  [1] "AircraftType"
##  [2] "AirportName"
##  [3] "Altitudebin"
##  [4] "AircraftMakeModel"
##  [5] "WildlifeNumberstruck"
##  [6] "EffectImpacttoflight"
##  [7] "EffectOther"
##  [8] "LocationNearbyifenroute"
##  [9] "AircraftFlightNumber"
## [10] "FlightDate"
## [11] "RecordID"
## [12] "EffectIndicatedDamage"
## [13] "LocationFreeformenroute"
## [14] "AircraftNumberofengines"
## [15] "AircraftAirlineOperator"
## [16] "OriginState"
## [17] "WhenPhaseofflight"
## [18] "ConditionsPrecipitation"
## [19] "Remainsofwildlifecollected"
## [20] "RemainsofwildlifesenttoSmithsonian"
## [21] "Remarks"
## [22] "ReportedDate"
## [23] "WildlifeSize"
## [24] "ConditionsSky"
## [25] "WildlifeSpecies"
## [26] "WhenTimeHHMM"
## [27] "WhenTimeofday"
## [28] "Pilotwarnedofbirdsorwildlife"
## [29] "CostAircrafttimeoutofservicehours"
## [30] "CostOtherinflationadj"
## [31] "CostRepairinflationadj"
## [32] "CostTotal"
## [33] "Milesfromairport"
## [34] "Feetaboveground"
## [35] "Numberofhumanfatalities"
## [36] "Numberofpeopleinjured"
## [37] "SpeedIASinknots"
```

```r
# Load MongoDB in R
library(mongolite)
```

```
## Warning: package 'mongolite' was built under R version 3.4.1
```

```r
# Load data into MongoDB
mongoBirdStrikes<- mongo("birdStrikes.df")
str(mongoBirdStrikes)
```

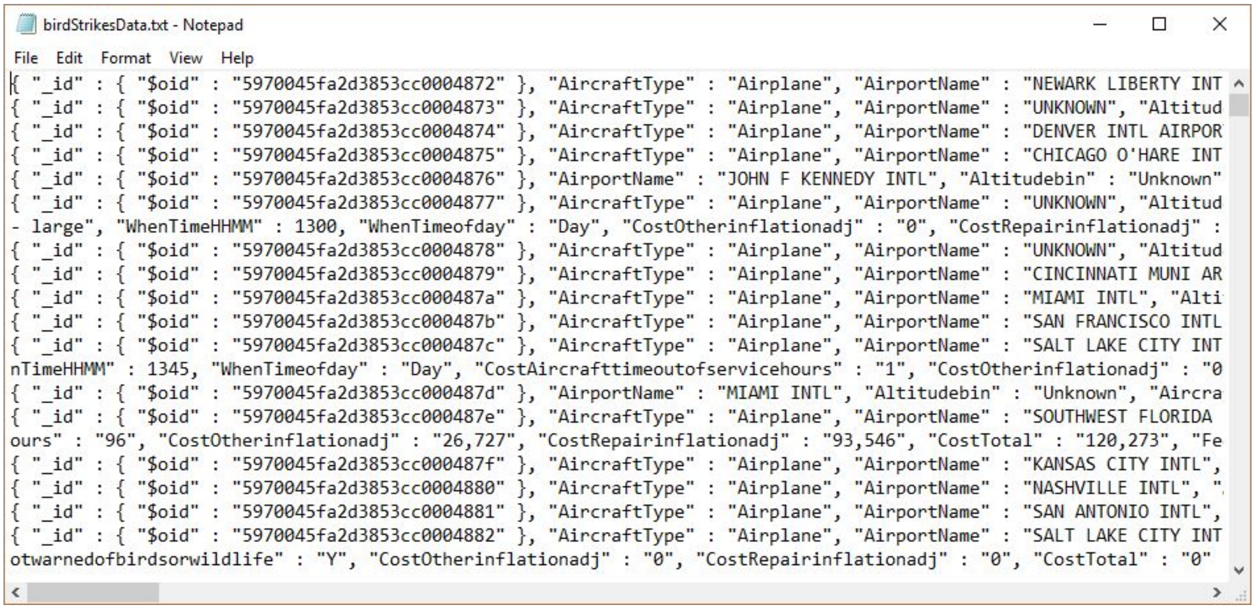
```
## Classes 'mongo', 'jeroen', 'environment' <environment: 0x00000000158a8678>
```

```r
# The insert function is used to insert data. The inserted data is a JSON object.
mongoBirdStrikes$insert(birdStrikes.df)
```

```
## List of 5
##  $ nInserted  : num 99404
##  $ nMatched   : num 0
##  $ nRemoved   : num 0
##  $ nUpserted  : num 0
##  $ writeErrors: list()
```

```r
# The inserted data can be viewed using the export function. Data is exported as a
# binary file.
mongoBirdStrikes$export(file("birdStrikesData.txt"))
```

Below is a screenshot of the exported data:



```
{ "_id" : { "$oid" : "5970045fa2d3853cc0004872" }, "AircraftType" : "Airplane", "AirportName" : "NEWARK LIBERTY INT
{ "_id" : { "$oid" : "5970045fa2d3853cc0004873" }, "AircraftType" : "Airplane", "AirportName" : "UNKNOWN", "Altitud
{ "_id" : { "$oid" : "5970045fa2d3853cc0004874" }, "AircraftType" : "Airplane", "AirportName" : "DENVER INTL AIRPOR
{ "_id" : { "$oid" : "5970045fa2d3853cc0004875" }, "AircraftType" : "Airplane", "AirportName" : "CHICAGO O'HARE INT
{ "_id" : { "$oid" : "5970045fa2d3853cc0004876" }, "AirportName" : "JOHN F KENNEDY INTL", "Altitudebin" : "Unknown"
{ "_id" : { "$oid" : "5970045fa2d3853cc0004877" }, "AircraftType" : "Airplane", "AirportName" : "UNKNOWN", "Altitud
- large", "WhenTimeHHMM" : 1300, "WhenTimeofday" : "Day", "CostOtherinflationadj" : "0", "CostRepairinflationadj" :
{ "_id" : { "$oid" : "5970045fa2d3853cc0004878" }, "AircraftType" : "Airplane", "AirportName" : "UNKNOWN", "Altitud
{ "_id" : { "$oid" : "5970045fa2d3853cc0004879" }, "AircraftType" : "Airplane", "AirportName" : "CINCINNATI MUNI AR
{ "_id" : { "$oid" : "5970045fa2d3853cc000487a" }, "AircraftType" : "Airplane", "AirportName" : "MIAMI INTL", "Alti
{ "_id" : { "$oid" : "5970045fa2d3853cc000487b" }, "AircraftType" : "Airplane", "AirportName" : "SAN FRANCISCO INTL
{ "_id" : { "$oid" : "5970045fa2d3853cc000487c" }, "AircraftType" : "Airplane", "AirportName" : "SALT LAKE CITY INT
nTimeHHMM" : 1345, "WhenTimeofday" : "Day", "CostAircrafttimeoutofservicehours" : "1", "CostOtherinflationadj" : "0
{ "_id" : { "$oid" : "5970045fa2d3853cc000487d" }, "AirportName" : "MIAMI INTL", "Altitudebin" : "Unknown", "Aircra
{ "_id" : { "$oid" : "5970045fa2d3853cc000487e" }, "AircraftType" : "Airplane", "AirportName" : "SOUTHWEST FLORIDA
ours" : "96", "CostOtherinflationadj" : "26,727", "CostRepairinflationadj" : "93,546", "CostTotal" : "120,273", "Fe
{ "_id" : { "$oid" : "5970045fa2d3853cc000487f" }, "AircraftType" : "Airplane", "AirportName" : "KANSAS CITY INTL",
{ "_id" : { "$oid" : "5970045fa2d3853cc0004880" }, "AircraftType" : "Airplane", "AirportName" : "NASHVILLE INTL", "
{ "_id" : { "$oid" : "5970045fa2d3853cc0004881" }, "AircraftType" : "Airplane", "AirportName" : "SAN ANTONIO INTL",
{ "_id" : { "$oid" : "5970045fa2d3853cc0004882" }, "AircraftType" : "Airplane", "AirportName" : "SALT LAKE CITY INT
otwarnedofbirdsorwildlife" : "Y", "CostOtherinflationadj" : "0", "CostRepairinflationadj" : "0", "CostTotal" : "0"
```

Task 3

Perform the following fetch operations:

a. Fetch the unique airport names from the database

I can us the distinct function to find unique instances of any particular column in the dataset.

```
querya<-mongoBirdStrikes$distinct("AirportName")

# This shows a summary of the results:
summary(querya)
```

```
##    Length     Class      Mode
##      1703 character character
```

```
# This shows the first six lines of the results:
head(querya)
```

```
## [1] "NEWARK LIBERTY INTL ARPT"          "UNKNOWN"
## [3] "DENVER INTL AIRPORT"               "CHICAGO O'HARE INTL ARPT"
## [5] "JOHN F KENNEDY INTL"               "CINCINNATI MUNI ARPT-LUNKEN FIELD"
```

b. Count the number of records where originState equals "New Jersey"

The count function is used to count the number of instances matching specific criteria:

```
queryb<-mongoBirdStrikes$count('{"OriginState":"New Jersey"}')

# This prints the result with context:
print(paste("The number of records where originState equals 'New Jersey' is", queryb))
```

```
## [1] "The number of records where originState equals 'New Jersey' is 2936"
```

c. Fetch the data with conditionsPrecipitation being fog and sort the data in descending order of recordId.

The find function can be used to find all the instances matching specific criteria. The sort function can sort the fetched data in ascending or descending order with value 1 for ascending and -1 for descending:

```
queryc<-mongoBirdStrikes$find('{"ConditionsPrecipitation":"Fog"}', sort='{"RecordID":-1}')

# This shows a summary of the results:
summary(queryc)
```

```
##  AircraftType       AirportName        Altitudebin
##  Length:878         Length:878         Length:878
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  AircraftMakeModel  WildlifeNumberstruck EffectImpacttoflight
##  Length:878         Length:878           Length:878
##  Class :character   Class :character     Class :character
##  Mode  :character   Mode  :character     Mode  :character
##
##
##
##
##   FlightDate           RecordID      EffectIndicatedDamage
##  Length:878         Min.   :  1207   Length:878
##  Class :character   1st Qu.:215896   Class :character
##  Mode  :character   Median :234617   Mode  :character
##                     Mean   :241752
##                     3rd Qu.:262812
##                     Max.   :321151
##
##  AircraftNumberofengines AircraftAirlineOperator OriginState
##  Length:878              Length:878              Length:878
##  Class :character        Class :character        Class :character
##  Mode  :character        Mode  :character        Mode  :character
##
##
##
##
##  WhenPhaseofflight  ConditionsPrecipitation Remainsofwildlifecollected
##  Length:878         Length:878              Mode :logical
##  Class :character   Class :character        FALSE:660
##  Mode  :character   Mode  :character        TRUE :218
##
##
##
##
##  RemainsofwildlifesenttoSmithsonian   Remarks          ReportedDate
##  Mode :logical                       Length:878         Length:878
##  FALSE:794                           Class :character   Class :character
##  TRUE :84                            Mode  :character   Mode  :character
```

```
##
##
##
##
##  WildlifeSize       ConditionsSky      WildlifeSpecies     WhenTimeHHMM
##  Length:878         Length:878         Length:878         Min.   :   0
##  Class :character   Class :character   Class :character   1st Qu.: 730
##  Mode  :character   Mode  :character   Mode  :character   Median : 910
##                                                           Mean   :1082
##                                                           3rd Qu.:1310
##                                                           Max.   :2345
##                                                           NA's   :231
##  WhenTimeofday      Pilotwarnedofbirdsorwildlife CostOtherinflationadj
##  Length:878         Length:878                   Length:878
##  Class :character   Class :character             Class :character
##  Mode  :character   Mode  :character             Mode  :character
##
##
##
##
##  CostRepairinflationadj  CostTotal         Feetaboveground
##  Length:878             Length:878         Length:878
##  Class :character       Class :character   Class :character
##  Mode  :character       Mode  :character   Mode  :character
##
##
##
##
##  Milesfromairport   SpeedIASinknots    AircraftFlightNumber
##  Length:878         Length:878         Length:878
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  CostAircrafttimeoutofservicehours LocationFreeformenroute
##  Length:878                        Length:878
##  Class :character                  Class :character
##  Mode  :character                  Mode  :character
##
##
##
##
##  EffectOther        LocationNearbyifenroute
##  Length:878         Length:878
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

```r
# This shows the first six lines of the results:
head(queryc)
```

```
##   AircraftType                  AirportName Altitudebin AircraftMakeModel
## 1     Airplane            REDDING MUNICIPAL   > 1000 ft           EMB-120
## 2     Airplane            HEATHROW - LONDON   < 1000 ft             A-330
## 3     Airplane             ADAMS FIELD ARPT   > 1000 ft          CL-RJ900
## 4     Airplane GEORGE BUSH INTERCONTINENTAL   < 1000 ft           EMB-145
## 5     Airplane          JOHN F KENNEDY INTL   < 1000 ft         B-737-800
## 6     Airplane              SACRAMENTO INTL     Unknown      CL-RJ100/200
##   WildlifeNumberstruck EffectImpacttoflight      FlightDate RecordID
## 1                    1                 None 12/30/2011 0:00   321151
## 2              2 to 10                 None  12/6/2010 0:00   320316
## 3                    1                 <NA> 12/13/2011 0:00   319957
## 4              2 to 10                 None 12/31/2011 0:00   319683
## 5            11 to 100                 None  12/5/2011 0:00   319537
## 6                    1                 None 11/29/2011 0:00   319476
##   EffectIndicatedDamage AircraftNumberofengines AircraftAirlineOperator
## 1             No damage                       2        SKYWEST AIRLINES
## 2             No damage                       2              US AIRWAYS
## 3             No damage                       2           MESA AIRLINES
## 4             No damage                       2      ATLANTIC SOUTHEAST
## 5             No damage                       2       AMERICAN AIRLINES
## 6             No damage                       2        SKYWEST AIRLINES
##   OriginState WhenPhaseofflight ConditionsPrecipitation
## 1  California          Approach                     Fog
## 2         N/A          Approach                     Fog
## 3    Arkansas          Approach                     Fog
## 4       Texas      Landing Roll                     Fog
## 5    New York      Take-off run                     Fog
## 6  California          Approach                     Fog
##   Remainsofwildlifecollected RemainsofwildlifesenttoSmithsonian
## 1                      FALSE                              FALSE
## 2                      FALSE                              FALSE
## 3                      FALSE                              FALSE
## 4                      FALSE                              FALSE
## 5                       TRUE                              FALSE
## 6                       TRUE                               TRUE
##
## 1
## 2
## 3
## 4
## 5 AT ROTATION, FLEW THRU A FLOCK WHICH WERE JUST LIFTING OFF. 10 PLUS STRIKES. NO DMG NOTED. 200 FT
## 6
##      ReportedDate WildlifeSize ConditionsSky       WildlifeSpecies
## 1 12/30/2011 0:00        Large      Overcast  Unknown bird - large
## 2            <NA>         <NA>          <NA>          Unknown bird
## 3            <NA>       Medium      Overcast Unknown bird - medium
## 4            <NA>        Small      Overcast  Unknown bird - small
## 5            <NA>        Small      Overcast          Snow bunting
## 6            <NA>        Small      Overcast           House finch
##   WhenTimeHHMM WhenTimeofday Pilotwarnedofbirdsorwildlife
## 1         1342           Day                            N
## 2         920          Day                         <NA>
## 3        1030          Day                            N
## 4          NA          Day                            Y
## 5        1400          Day                            N
## 6        1530          Day                            Y
##   CostOtherinflationadj CostRepairinflationadj CostTotal Feetaboveground
## 1                     0                      0         0           1,500
## 2                     0                      0         0              30
## 3                     0                      0         0           1,600
## 4                     0                      0         0               0
## 5                     0                      0         0               0
## 6                     0                      0         0            <NA>
##   Milesfromairport SpeedIASinknots AircraftFlightNumber
## 1             <NA>            <NA>                 <NA>
## 2                0             130                 <NA>
## 3             <NA>             175                 2681
## 4                0            <NA>                4672?
## 5                0             150                 1850
## 6             <NA>            <NA>                 4562
##   CostAircrafttimeoutofservicehours LocationFreeformenroute EffectOther
## 1                              <NA>                    <NA>        <NA>
## 2                              <NA>                    <NA>        <NA>
## 3                              <NA>                    <NA>        <NA>
## 4                              <NA>                    <NA>        <NA>
## 5                              <NA>                    <NA>        <NA>
## 6                              <NA>                    <NA>        <NA>
##   LocationNearbyifenroute
## 1                    <NA>
## 2                    <NA>
## 3                    <NA>
## 4                    <NA>
## 5                    <NA>
## 6                    <NA>
```

d. Fetch data.

Fetch only the following columns for aircraftAirlineOperator: "AMERICAN AIRLINES" and "CONTINENTAL AIRLINES"

1. recordId
2. originState
3. aircraftAirlineOperator
4. airportName

I began by breaking the find argument into two parts, query and fields. The query section is used as the search criteria. In order to find two things in the same column, I needed to use $in in the query section as shown below.

The fields section is used to specify the column(s) to display. To display any column write the column name and keep the value as 1. If a column does not need to be displayed, keep the value as 0. By default any column name not mentioned in field argument is not displayed. The column name _id is the default primary key for the record. To remove this column from the result just add another argument with value of _id 0.

```
queryd <- mongoBirdStrikes$find(
  query = '{"AircraftAirlineOperator" :
```

```
  { "$in" : [ "AMERICAN AIRLINES",
  "CONTINENTAL AIRLINES"]  } }',
  fields = '{"RecordID":1,"OriginState":1,
  "AircraftAirlineOperator":1,"AirportName":1, "_id":0}'
)

# This shows a summary of the results:
summary(queryd)
```

```
##  AirportName           RecordID      AircraftAirlineOperator
##  Length:4684        Min.   :200004   Length:4684
##  Class :character   1st Qu.:216298   Class :character
##  Mode  :character   Median :241124   Mode  :character
##                     Mean   :249953
##                     3rd Qu.:268905
##                     Max.   :319945
##  OriginState
##  Length:4684
##  Class :character
##  Mode  :character
##
##
##
```

```
# This shows the first six lines of the results:
head(queryd)
```

```
##                  AirportName RecordID AircraftAirlineOperator OriginState
## 1    NEWARK LIBERTY INTL ARPT   200508    CONTINENTAL AIRLINES  New Jersey
## 2                     UNKNOWN   204787       AMERICAN AIRLINES         N/A
## 3         MINETA SAN JOSE INTL   208470       AMERICAN AIRLINES  California
## 4      LAFAYETTE REGIONAL (LA)   204764    CONTINENTAL AIRLINES   Louisiana
## 5          JOHN F KENNEDY INTL   202568       AMERICAN AIRLINES    New York
## 6 DALLAS/FORT WORTH INTL ARPT   200470       AMERICAN AIRLINES       Texas
```