# DataFrames in R

### Loading and looking at your dataframe

The easiest data source to load is a csv. To read a csv into the global environment use the read_csv function. To see all the arguments of read_csv type ?read.csv in the console and run the command.

You can assign variables with either <- or the = sign. We name our data frame qbs below.

```
# NOTE: Make sure your .rmd file is in the same directory as your .csv file.
qbs <- read.csv('nfl_qbs.csv', stringsAsFactors = FALSE)
```

You can look at a preview of your data frame by double clicking on the data frame in the data UI to the left. If you want to look at it without the interface use the head() function.

```
head(qbs)
```

```
##         Player Year Age  Draft  Tm  Lg  G GS Cmp Att Cmp_per  Yds TD Int Pick6
## 1  Alex Tanney 2019  32        NYG NFL  1  0   1   1  100.00    1  0   0    NA
## 2    Tim Boyle 2019  25        GNB NFL  3  0   3   4   75.00   15  0   0    NA
## 3  Matt Schaub 2019  38 Mar-90 ATL NFL  6  1  50  67   74.63  580  3   1    NA
## 4   Drew Brees 2019  40 Feb-32 NOR NFL 11 11 281 378   74.34 2979 27   4    NA
## 5   Derek Carr 2019  28 Feb-36 OAK NFL 16 16 361 513   70.37 4054 21   8     2
## 6 Chase Daniel 2019  33        CHI NFL  3  1  45  64   70.31  435  3   2    NA
##   TD_per Int_per  Rate Sk Yds.1  Y.A AY.A ANY.A   Y.G  W  L  T X4QC GWD
## 1    0.0    0.00  79.2  0     0 1.00 1.00  1.00   1.0 NA NA NA   NA  NA
## 2    0.0    0.00  80.2  0     0 3.75 3.75  3.75   5.0 NA NA NA   NA  NA
## 3    4.5    1.49 109.0  2    19 8.66 8.88  8.35  96.7  0  1  0   NA  NA
## 4    7.1    1.06 116.3 12    89 7.88 8.83  8.33 270.8  8  3  0    1   2
## 5    4.1    1.56 100.8 29   184 7.90 8.02  7.25 253.4  7  9  0    2   3
## 6    4.7    3.13  91.6  7    48 6.80 6.33  5.03 145.0  0  1  0   NA  NA
```

You can view the first however many by adding an argument to head.

```
head(qbs, 3)
```

```
##        Player Year Age  Draft  Tm  Lg G GS Cmp Att Cmp_per Yds TD Int Pick6
## 1 Alex Tanney 2019  32        NYG NFL 1  0   1   1  100.00   1  0   0    NA
## 2   Tim Boyle 2019  25        GNB NFL 3  0   3   4   75.00  15  0   0    NA
## 3 Matt Schaub 2019  38 Mar-90 ATL NFL 6  1  50  67   74.63 580  3   1    NA
##   TD_per Int_per  Rate Sk Yds.1  Y.A AY.A ANY.A  Y.G  W  L  T X4QC GWD
## 1    0.0    0.00  79.2  0     0 1.00 1.00  1.00  1.0 NA NA NA   NA  NA
## 2    0.0    0.00  80.2  0     0 3.75 3.75  3.75  5.0 NA NA NA   NA  NA
## 3    4.5    1.49 109.0  2    19 8.66 8.88  8.35 96.7  0  1  0   NA  NA
```

Our data in this example has column titles and players in each row which help us identify the data. If you don't have this then you can name the columns and rows with colnames() and row.names(). The c() stands for concatenate, it creates a vector.

Notice I have eval=FALSE in the chunk options. This means we will not run the code even when we run the entire file. I also put a # in front, this makes the code a comment and no longer runnable.

```
#colnames(qbs) <- c("column1", "column2", "column3", "column4")

#row.names(qbs) <- c("player1", "player2", "player3", "player4")
```

To see the size of your data frame you can either look at the data UI or you can run dim(). The [] notation allows us to return the subset of the dim() vector. Vectors in R start with an index at 1 unlike python which starts at 0. Essentially, dim(qbs)[1] will return the first number in the dim(qbs) returned vector. You can also use nrow() and ncol() to get the dataframe size.

```
# Returns the number of rows and columns
dim(qbs)
```

```
## [1] 73 29
```

```
# number of rows
dim(qbs)[1]
```

```
## [1] 73
```

```
# number of columns
dim(qbs)[2]
```

```
## [1] 29
```

```
# number of rows
nrow(qbs)
```

```
## [1] 73
```

```
# number of columns
ncol(qbs)
```

```
## [1] 29
```

### Subsetting

You can subset data frames very easily with the [] notation. The subset is dataframe[rows,columns].

```
# returns the 1st row
qbs[1,]
```

```
##        Player Year Age Draft  Tm  Lg G GS Cmp Att Cmp_per Yds TD Int Pick6
## 1 Alex Tanney 2019  32       NYG NFL 1  0   1   1     100   1  0   0    NA
##   TD_per Int_per Rate Sk Yds.1 Y.A AY.A ANY.A Y.G  W  L  T X4QC GWD
## 1      0       0 79.2  0     0   1    1     1   1 NA NA NA   NA  NA
```

```
# returns the 1st column
qbs[,1]
```

```
##  [1] "Alex Tanney"        "Tim Boyle"          "Matt Schaub"
##  [4] "Drew Brees"         "Derek Carr"         "Chase Daniel"
##  [7] "Ryan Tannehill"     "Kirk Cousins"       "Jimmy Garoppolo"
## [10] "Teddy Bridgewater"  "Deshaun Watson"     "Colt McCoy"
## [13] "Tyrod Taylor"       "Matt Ryan"          "Russell Wilson"
## [16] "Lamar Jackson"      "Philip Rivers"      "Patrick Mahomes"
## [19] "Nick Foles"         "Joe Flacco"         "Dak Prescott"
## [22] "Matt Moore"         "Case Keenum"        "Kyler Murray"
## [25] "Luke Falk"          "Matthew Stafford"   "Drew Lock"
## [28] "Carson Wentz"       "Mitchell Trubisky"  "Jared Goff"
## [31] "Devlin Hodges"      "Mason Rudolph"      "Aaron Rodgers"
## [34] "Kyle Allen"         "Ryan Fitzpatrick"   "Sam Darnold"
## [37] "Eli Manning"        "Daniel Jones"       "Tom Brady"
## [40] "Jacoby Brissett"    "Jameis Winston"     "Gardner Minshew"
## [43] "Robert Griffin"     "Mike Glennon"       "Josh McCown"
## [46] "Andy Dalton"        "Marcus Mariota"     "Baker Mayfield"
## [49] "Jeff Driskel"       "Josh Allen"         "Dwayne Haskins"
## [52] "Sean Mannion"       "A.J. McCarron"      "Ben Roethlisberger"
## [55] "Cam Newton"         "David Blough"       "Will Grier"
## [58] "Brian Hoyer"        "Josh Rosen"         "Matt Barkley"
## [61] "Blake Bortles"      "Ryan Griffin"       "Taysom Hill"
## [64] "Trevor Siemian"     "Jarrett Stidham"    "Ryan Finley"
## [67] "Brandon Allen"      "Brett Hundley"      "Garrett Gilbert"
## [70] "David Fales"        "Trace McSorley"     "Nick Mullens"
## [73] "Cooper Rush"
```

```
# returns the value at the 15th row and first column
qbs[15, 1]
```

```
## [1] "Russell Wilson"
```

You can use combinations of c() and : to get more than single values!

```
# returns the 10th and 15th items in the first row
qbs[c(10, 15), 1]
```

```
## [1] "Teddy Bridgewater" "Russell Wilson"
```

```
# returns rows 10 through 15 and columns 1 through 3
qbs[10:15, 1:3]
```

```
##               Player Year Age
## 10 Teddy Bridgewater 2019  27
```

```
## 11    Deshaun Watson 2019  24
## 12        Colt McCoy 2019  33
## 13      Tyrod Taylor 2019  30
## 14         Matt Ryan 2019  34
## 15    Russell Wilson 2019  31
```

Another way to get a single vector column is with the $ sign.

```
qbs$Player
```

```
##  [1] "Alex Tanney"        "Tim Boyle"          "Matt Schaub"
##  [4] "Drew Brees"         "Derek Carr"         "Chase Daniel"
##  [7] "Ryan Tannehill"     "Kirk Cousins"       "Jimmy Garoppolo"
## [10] "Teddy Bridgewater"  "Deshaun Watson"     "Colt McCoy"
## [13] "Tyrod Taylor"       "Matt Ryan"          "Russell Wilson"
## [16] "Lamar Jackson"      "Philip Rivers"      "Patrick Mahomes"
## [19] "Nick Foles"         "Joe Flacco"         "Dak Prescott"
## [22] "Matt Moore"         "Case Keenum"        "Kyler Murray"
## [25] "Luke Falk"          "Matthew Stafford"   "Drew Lock"
## [28] "Carson Wentz"       "Mitchell Trubisky"  "Jared Goff"
## [31] "Devlin Hodges"      "Mason Rudolph"      "Aaron Rodgers"
## [34] "Kyle Allen"         "Ryan Fitzpatrick"   "Sam Darnold"
## [37] "Eli Manning"        "Daniel Jones"       "Tom Brady"
## [40] "Jacoby Brissett"    "Jameis Winston"     "Gardner Minshew"
## [43] "Robert Griffin"     "Mike Glennon"       "Josh McCown"
## [46] "Andy Dalton"        "Marcus Mariota"     "Baker Mayfield"
## [49] "Jeff Driskel"       "Josh Allen"         "Dwayne Haskins"
## [52] "Sean Mannion"       "A.J. McCarron"      "Ben Roethlisberger"
## [55] "Cam Newton"         "David Blough"       "Will Grier"
## [58] "Brian Hoyer"        "Josh Rosen"         "Matt Barkley"
## [61] "Blake Bortles"      "Ryan Griffin"       "Taysom Hill"
## [64] "Trevor Siemian"     "Jarrett Stidham"    "Ryan Finley"
## [67] "Brandon Allen"      "Brett Hundley"      "Garrett Gilbert"
## [70] "David Fales"        "Trace McSorley"     "Nick Mullens"
## [73] "Cooper Rush"
```

Notice that when you return a single column the output is a vector but when it is more than one you return a data frame.

```
qbs[, c('Player', 'Age')]
```

```
##                Player Age
## 1         Alex Tanney  32
## 2           Tim Boyle  25
## 3         Matt Schaub  38
## 4          Drew Brees  40
## 5          Derek Carr  28
## 6        Chase Daniel  33
## 7      Ryan Tannehill  31
## 8        Kirk Cousins  31
## 9     Jimmy Garoppolo  28
## 10 Teddy Bridgewater  27
## 11    Deshaun Watson  24
## 12        Colt McCoy  33
## 13       Tyrod Taylor  30
## 14          Matt Ryan  34
```

```
## 15     Russell Wilson  31
## 16      Lamar Jackson  22
## 17      Philip Rivers  38
## 18    Patrick Mahomes  24
## 19         Nick Foles  30
## 20         Joe Flacco  34
## 21       Dak Prescott  26
## 22         Matt Moore  35
## 23        Case Keenum  31
## 24       Kyler Murray  22
## 25          Luke Falk  25
## 26   Matthew Stafford  31
## 27          Drew Lock  23
## 28       Carson Wentz  27
## 29  Mitchell Trubisky  25
## 30         Jared Goff  25
## 31      Devlin Hodges  23
## 32      Mason Rudolph  24
## 33      Aaron Rodgers  36
## 34         Kyle Allen  23
## 35   Ryan Fitzpatrick  37
## 36        Sam Darnold  22
## 37        Eli Manning  38
## 38       Daniel Jones  22
## 39          Tom Brady  42
## 40    Jacoby Brissett  27
## 41     Jameis Winston  25
## 42    Gardner Minshew  23
## 43     Robert Griffin  29
## 44       Mike Glennon  30
## 45        Josh McCown  40
## 46        Andy Dalton  32
## 47     Marcus Mariota  26
## 48     Baker Mayfield  24
## 49       Jeff Driskel  26
## 50         Josh Allen  23
## 51     Dwayne Haskins  22
## 52       Sean Mannion  27
## 53      A.J. McCarron  29
## 54 Ben Roethlisberger  37
## 55         Cam Newton  30
## 56       David Blough  24
## 57         Will Grier  24
## 58        Brian Hoyer  34
## 59         Josh Rosen  22
## 60       Matt Barkley  29
## 61      Blake Bortles  27
## 62       Ryan Griffin  30
## 63        Taysom Hill  29
## 64     Trevor Siemian  28
## 65    Jarrett Stidham  23
## 66        Ryan Finley  25
## 67      Brandon Allen  27
## 68      Brett Hundley  26
```

```
## 69    Garrett Gilbert  28
## 70        David Fales  29
## 71     Trace McSorley  24
## 72       Nick Mullens  24
## 73        Cooper Rush  26
```

You can also subset the dataframe by looking for certain values.

```
qbs[qbs$Player == "Jimmy Garoppolo", ]
```

```
##            Player Year Age  Draft  Tm  Lg  G GS Cmp Att Cmp_per  Yds TD Int
## 9 Jimmy Garoppolo 2019  28 Feb-62 SFO NFL 16 16 329 476   69.12 3978 27  13
##   Pick6 TD_per Int_per Rate Sk Yds.1  Y.A AY.A ANY.A   Y.G  W L T X4QC GWD
## 9     1    5.7    2.73  102 36   237 8.36 8.26  7.22 248.6 13 3 0    4   4
```

### Computing in dataframes

It is pretty easy to compute values from the dataframe Generally, you want to compute values from columns. Below we return the mean and median of TDs from the QBs.

```
mean(qbs$TD)
```

```
## [1] 10.83562
```

```
median(qbs$TD)
```

```
## [1] 6
```

You can get your basic distribution with the summary() function.

```
summary(qbs$TD)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    6.00   10.84   21.00   36.00
```

Now, with what we learned before, what if we wanted all the information of qbs who threw more than 30 TDs?

```
qbs[qbs$TD > 30, ]
```

```
##            Player Year Age  Draft  Tm  Lg  G GS Cmp Att Cmp_per  Yds TD Int
## 15 Russell Wilson 2019  31 Mar-75 SEA NFL 16 16 341 516   66.09 4110 31   5
## 16  Lamar Jackson 2019  22 Jan-32 BAL NFL 15 15 265 401   66.08 3127 36   6
## 41 Jameis Winston 2019  25  1-Jan TAM NFL 16 16 380 626   60.70 5109 33  30
##    Pick6 TD_per Int_per  Rate Sk Yds.1  Y.A AY.A ANY.A   Y.G  W L T X4QC GWD
## 15     2    6.0    0.97 106.3 48   319 7.97 8.73  7.42 256.9 11 5 0    4   5
## 16    NA    9.0    1.50 113.3 23   106 7.80 8.92  8.19 208.5 13 2 0    1   2
## 41     7    5.3    4.79  84.3 47   282 8.16 7.06  6.15 319.3  7 9 0    2   2
```

or we can use the which function

```
qbs[which(qbs$TD > 30), ]
```

```
##            Player Year Age  Draft  Tm  Lg  G GS Cmp Att Cmp_per  Yds TD Int
## 15 Russell Wilson 2019  31 Mar-75 SEA NFL 16 16 341 516   66.09 4110 31   5
## 16  Lamar Jackson 2019  22 Jan-32 BAL NFL 15 15 265 401   66.08 3127 36   6
## 41 Jameis Winston 2019  25  1-Jan TAM NFL 16 16 380 626   60.70 5109 33  30
##    Pick6 TD_per Int_per  Rate Sk Yds.1  Y.A AY.A ANY.A   Y.G  W L T X4QC GWD
## 15     2    6.0    0.97 106.3 48   319 7.97 8.73  7.42 256.9 11 5 0    4   5
## 16    NA    9.0    1.50 113.3 23   106 7.80 8.92  8.19 208.5 13 2 0    1   2
## 41     7    5.3    4.79  84.3 47   282 8.16 7.06  6.15 319.3  7 9 0    2   2
```

The which function returns an index value and the subset uses those index values to decide what to keep.

Which QB has the most TDs? Notice we use ==. = assigns variables while == is what we think of when we say equals.

```
qbs$Player[which(qbs$TD == max(qbs$TD))]
```

```
## [1] "Lamar Jackson"
```

or to get the player and value

```
qbs[which(qbs$TD == max(qbs$TD)), c('Player', 'TD')]
```

```
##           Player TD
## 16 Lamar Jackson 36
```

### Missing Data

To check if our data is missing values we use the is.na function. Combined with the any you can see if your data frame has missing data.

```
any(is.na(qbs))
```

```
## [1] TRUE
```

let's remove the na's with na.omit

```
qbs_com <- na.omit(qbs)

dim(qbs_com)[1]
```

```
## [1] 14
```

Now we see only 14 QBs had all columns filled. That's not even the entire leagues starting QBs. Why is that? Be careful with data. In this example, one of the reasons is Pick6's. A Pick6 is recorded only when it happens so instead of a zero when a QB didn't throw any in the season it is missing. Instead let's make our new data frame starter QBs who threw at least 200 passes.

```
qbs_start <- qbs[(qbs$GS > 0 & qbs$Att >= 200), ]
```

Now we are down to 33 QBs.

Lastly, we can create new columns from others!

```
qbs_start$td_int_ratio <- qbs_start$TD_per/qbs_start$Int_per
```

### Practice - Class Activity

Create a new data frame called mvp. Start by further narrowing the qbs_start data frame by only keeping the top half of QBs by the number of TDs thrown. That means greater than or equal to the top 50 percent.

How many QBs are in the discussion?

Show the distribution of the TDs by these QBs

Using what you learned show the top three QBs in total number of TDs. (Hint: Look up the order function)

Now, I like a QB who is efficient. Let's narrow it down to 5 candidates by taking the top 5 QBs in td_int_ratio!

Finally, we will decide our MVP based on a weighted scale. Create a column called rank that is the sum of 20 percent times cmp_per (completion percentage), 30 percent times TD, 30 percent times td_int_ratio, and 20 percent times GWD (game winning drives). Show the QBs, their statistics used to create rank, and rank in the final output dataframe. The highest overall rank value is our MVP.

Turn your R-notebook into the appropriate Gradescope Assignment. Turn it in by the end of class