

# Classwork

Immanuel Williams Ph.D.

8/17/2020

**Every Time, you create a dataframe, explore it**

**Monday: Review & Compose a R Markdown File**

**Your work should be done a R Markdown file**

## 1. Extraction (Do work in an R Markdown)

Extract NBA Data with MJ in 1993 from NBA Reference:

1. View website in browser:  
<https://www.basketballreference.com/players/j/jordami01/gamelog/1993/>
  - Hover over column to determine column name meanings.
2. Extract using code

```
library(tidyverse) ## data manipulations and visualization
library(XML) ## Extract urls
library(httr) ## Extract urls
rm(list = ls())

## Specify website/url
url = "https://www.basketball-reference.com/players/j/jordami01/gamelog/1991/"

## Open website
get_url = GET(url)

## Get Tables
tab_url = readHTMLTable(rawToChar(get_url$content), stringsAsFactors = F)

## Extract main dataframe
mj_1991_df = tab_url$pgl_basic
```

3. View Dataframe in R

**Q1.** Specify the variable types for all the variables. What R function should be use to see all the variables?

## 2. Manipulation, Learning how to access information in dataframe

### Accessing certain aspects of the dataframe using brackets

#### Getting access to rows and columns at the same time

```
## df[row,column]
mtcars[1,2]

## [1] 6
```

#### Getting access to just rows

```
## df[row,]
mtcars[c(1,2),]

##           mpg cyl disp  hp drat   wt  qsec vs am gear carb
## Mazda RX4      21   6  160 110  3.9 2.620 16.46  0  1    4    4
## Mazda RX4 Wag  21   6  160 110  3.9 2.875 17.02  0  1    4    4
```

#### Getting access to columns

```
## df[,column]
mtcars[,c(1,2)]

##           mpg cyl
## Mazda RX4      21.0   6
## Mazda RX4 Wag  21.0   6
## Datsun 710      22.8   4
## Hornet 4 Drive  21.4   6
## Hornet Sportabout 18.7   8
## Valiant         18.1   6
## Duster 360      14.3   8
## Merc 240D       24.4   4
## Volvo 142E      21.4   4
```

### Methods to remove columns (variables) from a dataframe

*Method 1:* The brackets function can be used to extract specific column based on the column number. Horse power is the 4th variable in the mtcars dataframe so in order to remove it we will use the minus sign '-' and the number 4.

```
mtcars %>%
  head()

##           mpg cyl disp  hp drat   wt  qsec vs am gear carb
## Mazda RX4      21.0   6  160 110  3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag  21.0   6  160 110  3.90 2.875 17.02  0  1    4    4
## Datsun 710      22.8   4  108  93  3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive  21.4   6  258 110  3.08 3.215 19.44  1  0    3    1
```

```
## Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2
## Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1
```

```
mtcars[, -4] %>%
  head()
```

```
##           mpg cyl disp drat   wt  qsec vs am gear carb
## Mazda RX4      21.0   6  160 3.90 2.620 16.46  0  1   4    4
## Mazda RX4 Wag  21.0   6  160 3.90 2.875 17.02  0  1   4    4
## Datsun 710     22.8   4  108 3.85 2.320 18.61  1  1   4    1
## Hornet 4 Drive  21.4   6  258 3.08 3.215 19.44  1  0   3    1
## Hornet Sportabout 18.7   8  360 3.15 3.440 17.02  0  0   3    2
## Valiant        18.1   6  225 2.76 3.460 20.22  1  0   3    1
```

*Method 2:* The `select()` function and the minus sign '-' can be used to remove not needed variables. Below is an example of removing the horse power 'hp' from *mtcars*.

```
## With hp
mtcars %>%
  head()
```

```
##           mpg cyl disp  hp drat   wt  qsec vs am gear carb
## Mazda RX4      21.0   6  160 110 3.90 2.620 16.46  0  1   4    4
## Mazda RX4 Wag  21.0   6  160 110 3.90 2.875 17.02  0  1   4    4
## Datsun 710     22.8   4  108  93 3.85 2.320 18.61  1  1   4    1
## Hornet 4 Drive  21.4   6  258 110 3.08 3.215 19.44  1  0   3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0   3    2
## Valiant        18.1   6  225 105 2.76 3.460 20.22  1  0   3    1
```

```
## Without hp
mtcars %>%
  select(-hp) %>%
  head()
```

```
##           mpg cyl disp drat   wt  qsec vs am gear carb
## Mazda RX4      21.0   6  160 3.90 2.620 16.46  0  1   4    4
## Mazda RX4 Wag  21.0   6  160 3.90 2.875 17.02  0  1   4    4
## Datsun 710     22.8   4  108 3.85 2.320 18.61  1  1   4    1
## Hornet 4 Drive  21.4   6  258 3.08 3.215 19.44  1  0   3    1
## Hornet Sportabout 18.7   8  360 3.15 3.440 17.02  0  0   3    2
## Valiant        18.1   6  225 2.76 3.460 20.22  1  0   3    1
```

**Q2. What variables do not have proper names in `mj_1991_df`?**

## Change column names (This is one way)

We can change the column name's (or variable name) of a dataframe by using the `colnames()` function, the column's index that I want to change, and the new name. In the example below I am going to change 'hp' to 'Horse\_Power'. I first find the index by running `colnames(mtcars)`.

```
## To find the index
colnames(mtcars)

## [1] "mpg"  "cyl"  "disp" "hp"    "drat" "wt"    "qsec" "vs"    "am"    "gear"
## [11] "carb"
```

```
## Make Change
colnames(mtcars)[4] = 'Horse_Power'
```

```
## Make sure it changed
head(mtcars)
```

	mpg	cyl	disp	Horse_Power	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	

**Q3.** The variables that seem to be questionable, change their name to be *Game\_Location* and *Game\_Outcome* based on the information in that column, in the **mj\_1991\_df**.

**Q4.** Within the *Game\_Location* variable, change the rows that have the '@' symbol to be 'Away' and the ' ' to be 'Home', in the **mj\_1991\_df**. Save the dataframe with a **mod1\_** in front.

## Remove Rows from a dataframe

We can use the filter function to only obtain rows that meet a specific requirement

```
mtcars %>%
  filter(vs!=0) %>%
  head()
```

	mpg	cyl	disp	Horse_Power	drat	wt	qsec	vs	am	gear	carb
## Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
## Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
## Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
## Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
## Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
## Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4

**Q5.** Look through **mod1\_mj\_1991\_df**. There are rows that are unnecessary, remove them from the dataframe. Save the dataframe with a **mod2\_** in front.

**Q6.** How many away and home games did MJ play that year?

**Q7.** The Game Score (GmSc) variable indicates the productivity of a player during a game. Create a new variable name *Binary\_GmSc* that distinguishes between above average named *Above\_Average\_GmSc* and below or equal to the average named *Below\_Average\_GmSc* based on the game score variable, in the **mod2\_mj\_1991\_df**. Save the dataframe with a **mod3\_** in front. Use the mean() function to find the average. An error is going to come up, problem solve this issue.

## Checking Variable Type

Use the class variable to check type of Variable

```
class(mod3_mj_1991_df$GmSc)
## [1] "numeric"
class(mod3_mj_1991_df$Game_Location)
## [1] "character"
class(mod3_mj_1991_df$PTS)
## [1] "character"
```

**Q8.** Fix the following variables based on the error found in question 7:

- FG, FGA, `3P`, `3PA`, ORB, DRB, AST, STL, BLK, PTS

Use **mod3\_mj\_1991\_df**. Save the dataframe with a **mod4\_** in front. For the variables that have a number in front of the variable, you must use tick marks around it. For example, ``3P` = as.numeric(`3P`)`. You can change the problematic variables if you like.

### 3. Visualizations

**Q9.** Create a plot to visualize the relationship between *Game\_Location* and *PTS* (Points). Use the **mod4\_mj\_1991\_df**. Make sure to label the x and y axes as well as the title. Make your plot look nice by changing the theme. Make your plot look nice by changing the theme. What pattern do you notice in this plot?

**Q10.** Create a plot to visualize the relationship between *AST* (Assists) and *PTS* (Points). Use the **mod4\_mj\_1991\_df**. Facet this relationship on *Game\_Location*. Make sure to label the x and y axes as well as the title. Make sure the title is centered. Make your plot look nice by changing the theme. What pattern do you notice in this plot?

**Q11.** Create plot using one quantitative and two categorical variables. Use variables that are different from **Q9.** and **Q10.** and make sense in the context in basketball. Make sure to label the x and y axes as well as the title. Make sure the title is centered. What pattern do you notice in this plot?