# Shared note-taking document for Jan 31, 2020 R workshop

Feel free to add anything useful!

Wrangling: more columns or rows than you need--narrow it down to what you want.
Cleaning: typos, garbage, find it and then decide what you want to do with it.
Create new functions to avoid repeating codes

In RStudio:
Assign a variable with <- or =, general practice is to use <-
Text variables need to be in quotes ‘Dan’
Variables are case sensitive
Pull up help with ?typeof or any other function
Use == to ask question
“
“Logical” represents trues and falses, “character” is for words (?)

! means not
| means or
# start a comment with hashtag

Console allows you to run one code at a time and run
Use R Script instead and save to go back
Can code in R Script and then run each line using short cut from Code, run selected lines or use short cuts; Run icon in R Script will run entire code

To clear data, click the broom icon in the top right box (under environment)

Function
Use c(x,y,z) to compose a vector (list of variables)

Use the escape key (ESC) if you need to get your > prompt back in the console window and R isn’t cooperating otherwise

Rm means to remove
Df means data frame - creating a data set

Transfering variables example - house_df$ward <- factor(house_df$ward)
Asking for data frame
house_df['price']
  price

1 190000
2 210000
3 143500
4 900000
5 900001
Asking for values
> house_df$price
[1] 190000 210000 143500 900000 900001

Parenthesis use with a function
Bracket use with matrix or dataframe

Factor means treating variable as category
> summary(gap_df) to get summary of data file, replace gap_df with data file name

y<-x+2
last.name<- 'Kerchner'
word<-10
word==7
last.name=='Washington'
house.price<-260000
house.price>200000
(word==7)& (house.price<=200000)

Degrees <- c('BA', 'BS', 'MS', 'PHD', 'MD', 'JD')
Tuition <- c(500, 1000, 750)
Mixed <- c('X', 6, TRUE)
mixed<- c(TRUE, 5, False, 2)

mixed<- c(TRUE, "asdfasuf", "aufsda")

Degrees[1]
Degrees[3]
Degrees[2:4]
Degrees[100]
Degrees[-1] all except the first variable
x<- -12
Degrees[3] <- 'MA'
Rm  to remove a specific value

```
Tuition *0.9
Tuition > 600
Discounted.tuition <- tuition * 0.9
High.tuition <- tuition > 600
Tuition[2]
tuition[c(TRUE, FALSE, TRUE]
Tuition[ high.tuition ]
Height <- c(72, 71, 60, 61, 59, 59, 57, 70, 74, 74, 73,
        69, 89, 81, 72, 70, 71, 69, 67, 59)
hist(height)
boxplot(height)
?boxplot
```

**Starting a new R script at 10:59am**

```
Name the script dataframe.R
In the environment pane, hit the broom icon
price <- c(190000, 210000, 143500, 90000, 900000)
ward <- c(7, 10, 3, 1, 1)
type <- c('A', 'H', 'H', 'A', 'H')

house_df <- data.frame(price, ward, type)
View(house_df)

typeof(house_df)
house_df$ward
house_df$price
house_df$ward <- factor(house_df$ward)
house_df$price
house_df [2,3]
house_df[2, 'type']
house_df[2, ]
house_df[2:4, ]
house_df[2:4, c('price', 'type')]
        The 2:4 is the rows you want, then the words indicate which column in this example


Gap_df <- read.csv('http://go.gwu.edu/gapminder')

install.packages("dplyr")
library(dplyr)
help(dplyr)
Dplyr is a package useful for wrangling data.
```

The pipe“%>%”

Means to run something on the left through something else on the right of the %>%

```
gap_df %>% select(country, gdpPercap)
```

Piping is like filtering but need dplyr (or applying a function to whatever datafile you are using)
Filter is choosing rows for test
Select to choose certain columns

```
gap_df2 <- gap_df %>% select(country, gdpPercap, year) %>%
filter(year==2007)


write.csv (“gap_df3.csv”)
```

== running a test
= is stating what it is

Github - can hook up R with Github
Different projects use different libraries

```
ggplot(data = gap_df) + aes(y = lifeExp, x = gdpPercap)+
  geom_point(alpha=0.1)
# the alpha here is for the transparency of the points themselves, within ggplot package
```

scale_x_log10() - scale x by 10

```
gap_df2007$loggdp<-log10(gdp_df2007$gdpPercap)
View(gdp_df2007)

gap.lm <- lm(data = gap_df2007,
        formula = lifeExp ~ loggdp + year + continent)

gap.lm <- lm(data = gap_df2007,
        formula = lifeExp ~ loggdp + continent)

summary(gap.lm)

coefficients(gap.lm)

coefficients(gap.lm)[‘loggdp’]
```

Close the project