DATA VISUALISATION WITH R

DR. ALVIN ANG

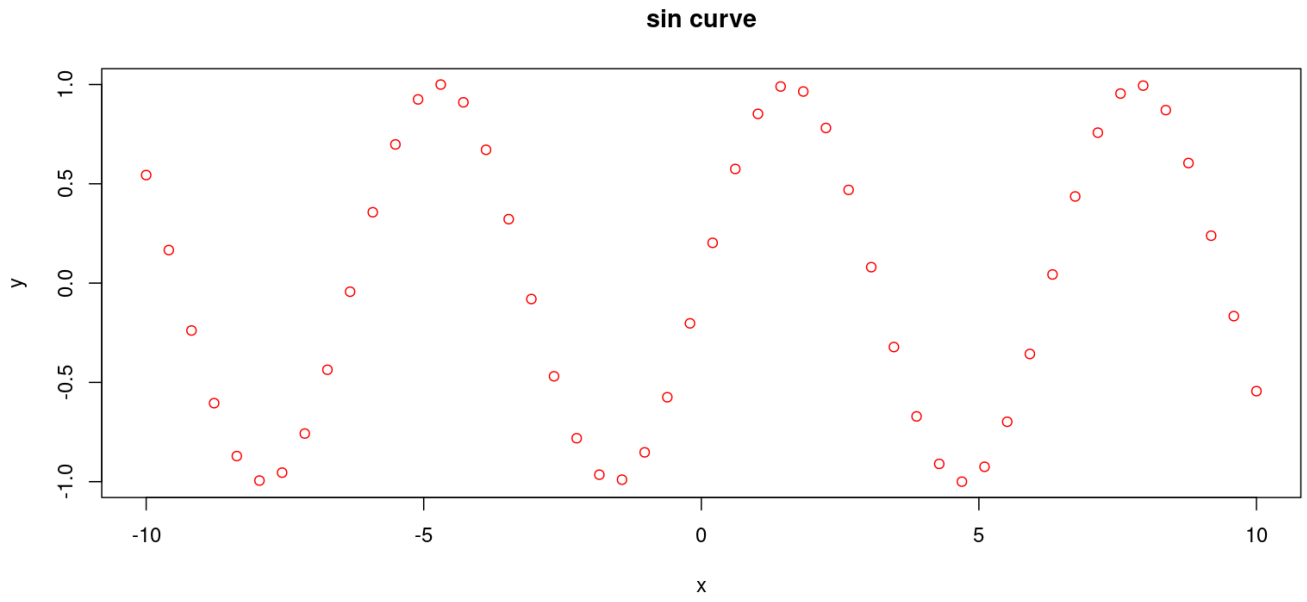
CONTENTS

I. USING BASIC R TO PLOT

FILE: https://www.alvinang.sg/s/Data-Visualisation-with-BASIC-R-by-Dr-Alvin-Ang.R

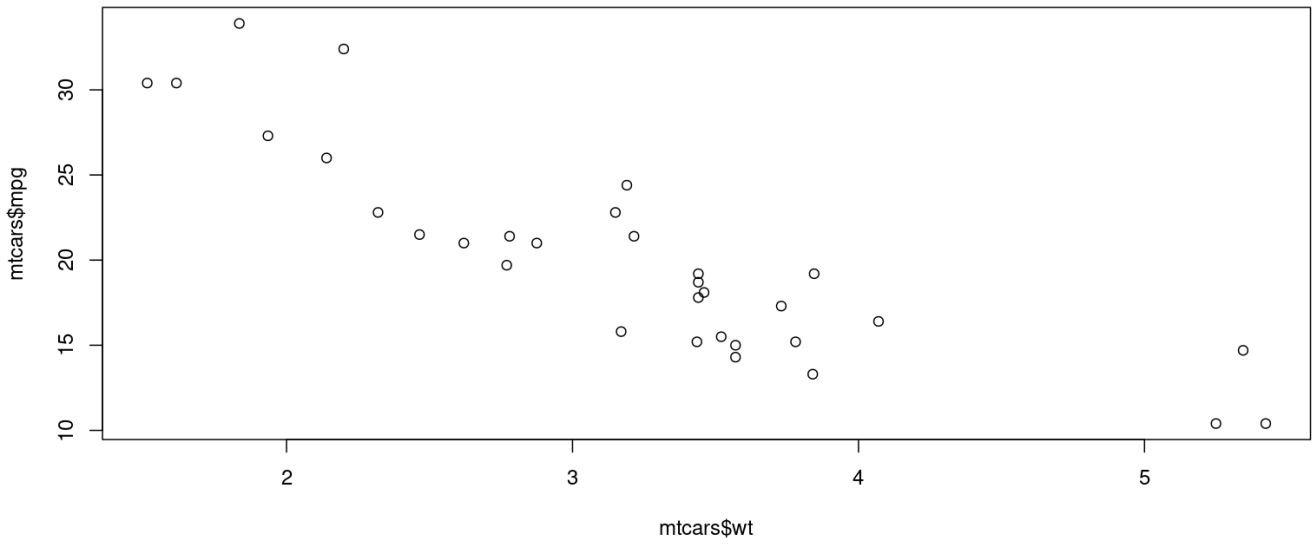
A. SCATTER PLOT 1

```
#a) Scatter Plot 1

x <- seq(-10,10,length.out=50)
y <- sin(x)
plot(x,y,main='sin curve',xlab='x',ylab='y',col='red')
```



B. SCATTER PLOT 2

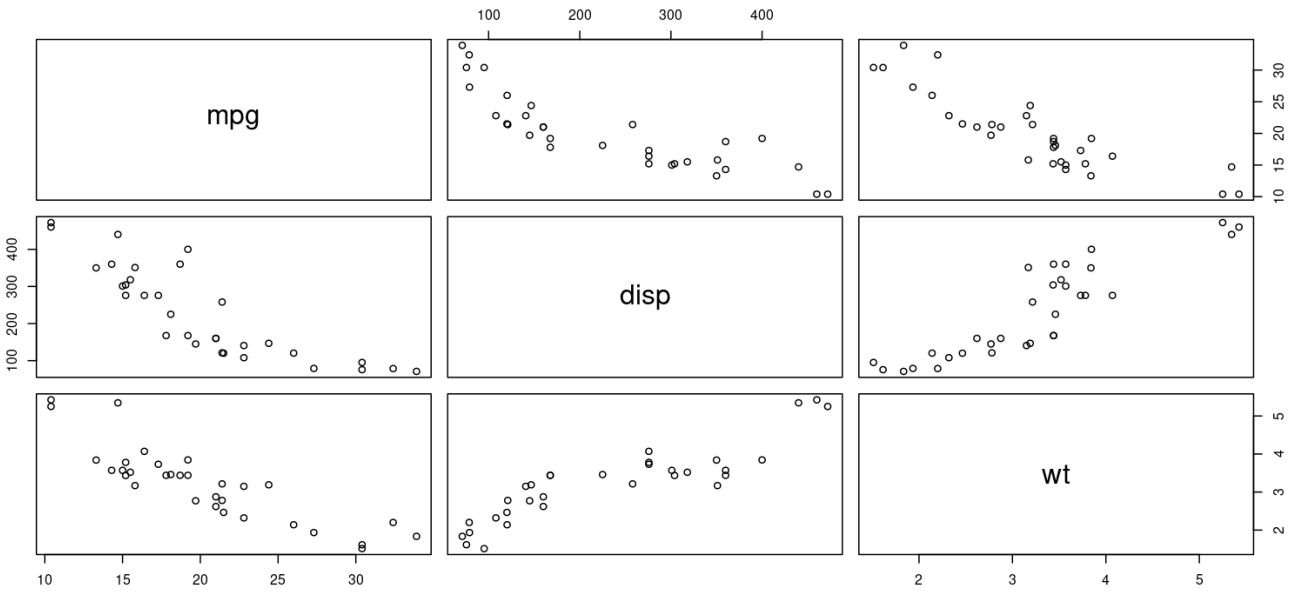
```
#b) Scatter Plot 2

plot(mtcars$wt, mtcars$mpg)
```



C. SCATTER PLOT 3

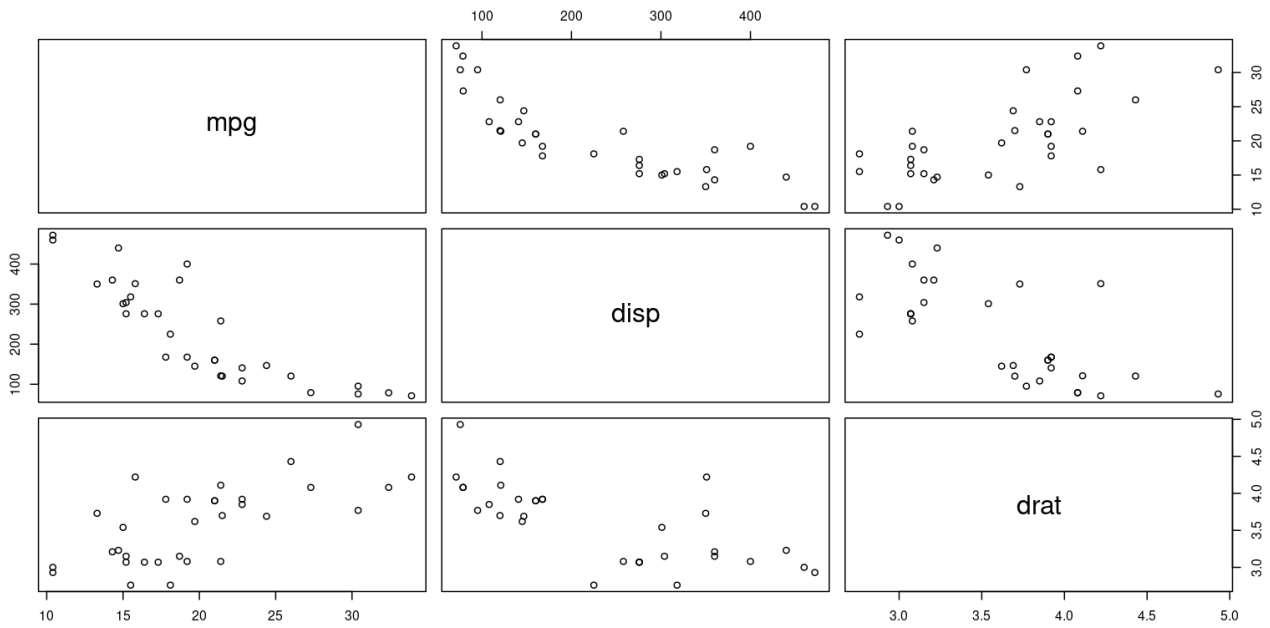
1. PAIR PLOT 1

```
#c)(i) Pair Plot 1
plot(mtcars[c('mpg','disp','wt')])
```



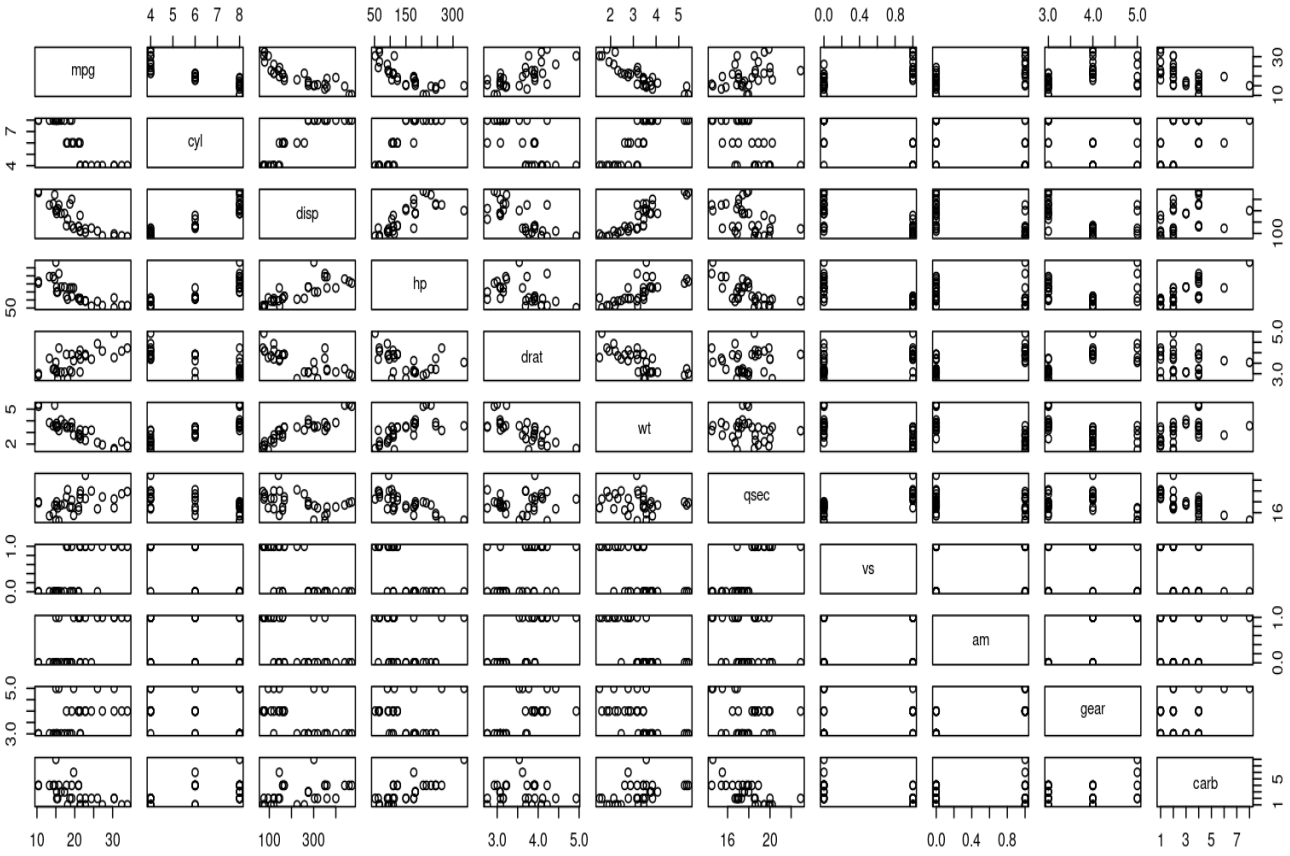
2. PAIR PLOT 2

```
#c)(ii) Pair Plot 2
plot(mtcars[c(1,3,5)])
```



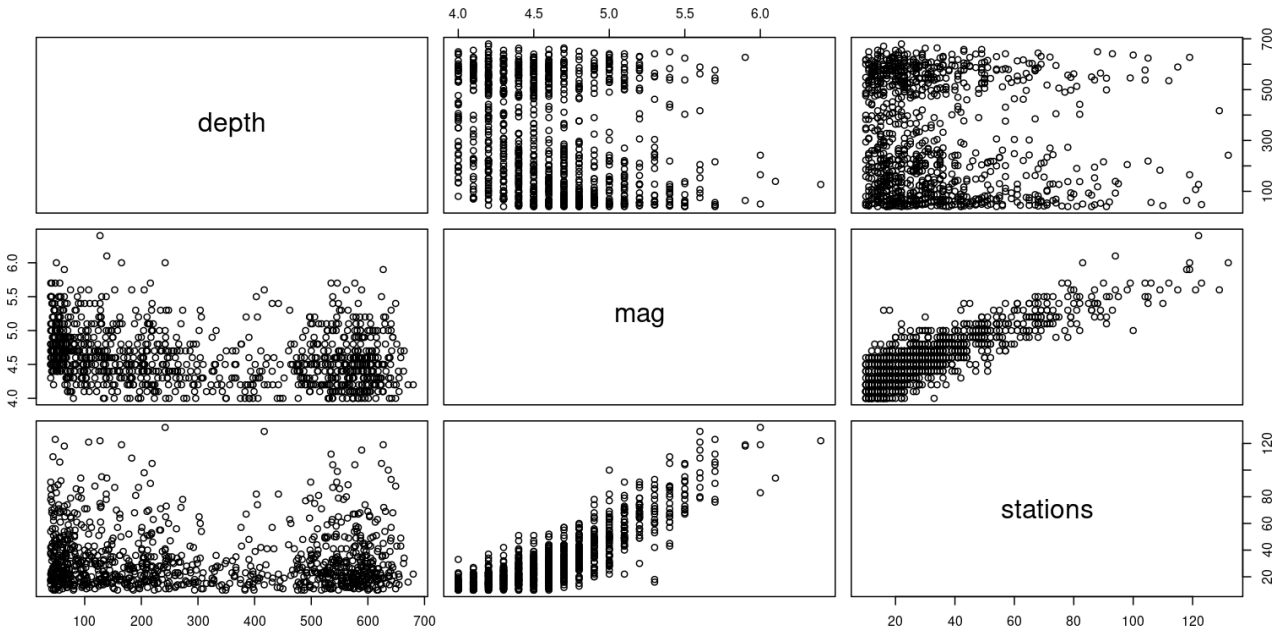
3. PAIR PLOT 3

```
#c)(iii) Pair Plot 3
plot(mtcars)
```



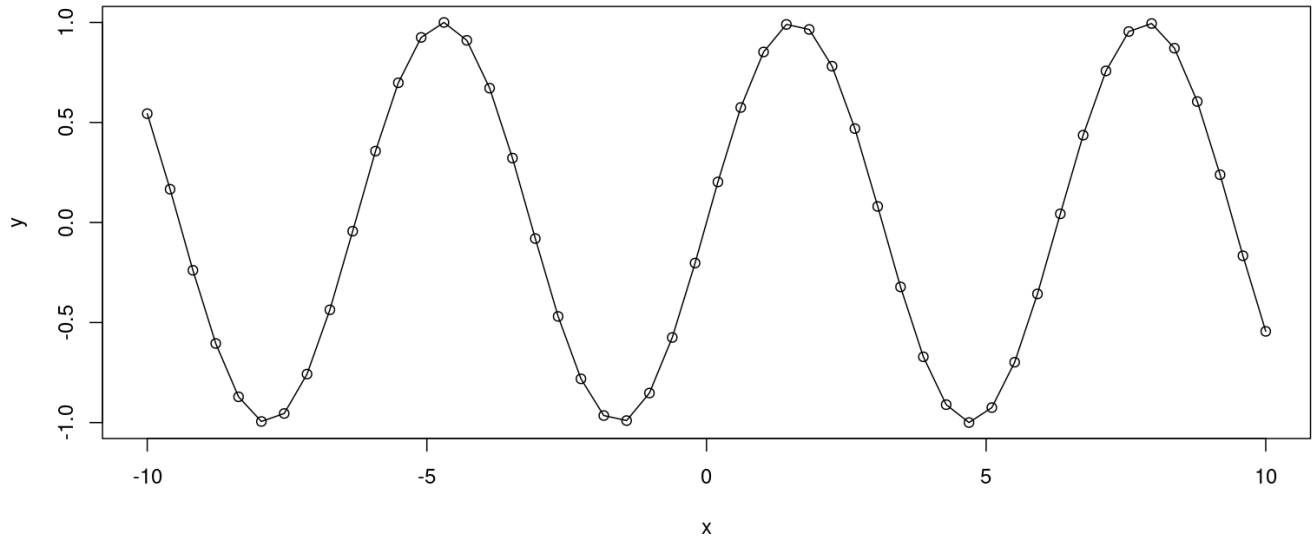
4. PAIR PLOT 4

```
#c)(iv) Pair Plot 4
plot(quakes[c('depth','mag','stations')])
```



D. LINE PLOT

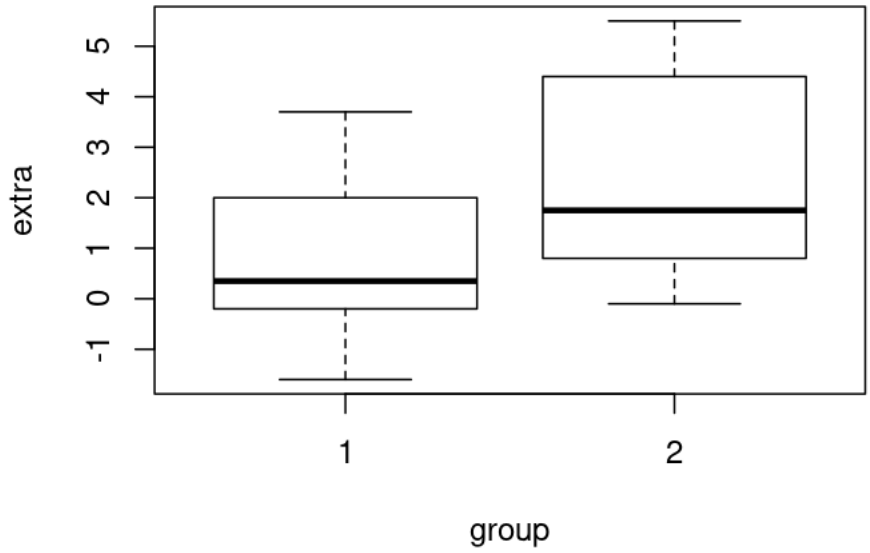
```
#d) Line Plot
x <- seq(-10,10,length.out=50)
y <- sin(x)
plot(x,y)
lines(x,y)
```



E. BOX PLOT

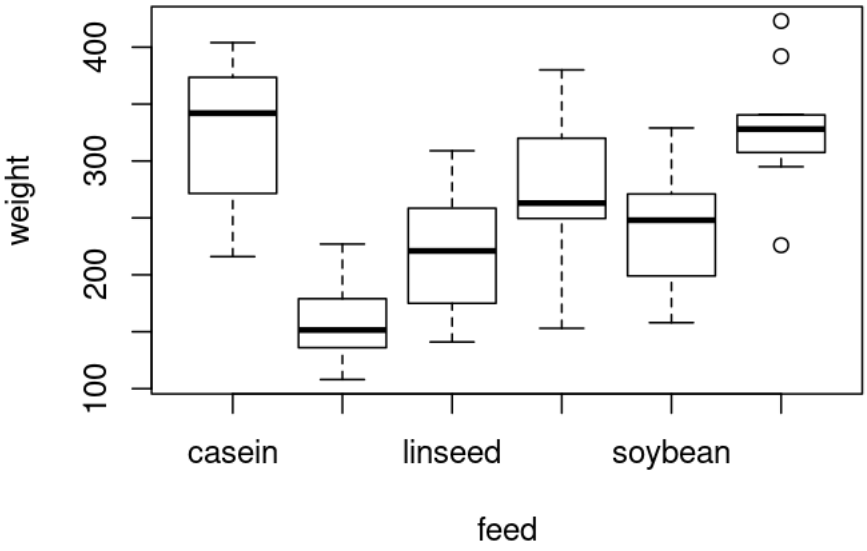
1. BOX PLOT 1

```
#e)(i) Box Plot 1
boxplot(extra~group,data=sleep)
```



2. BOX PLOT 2

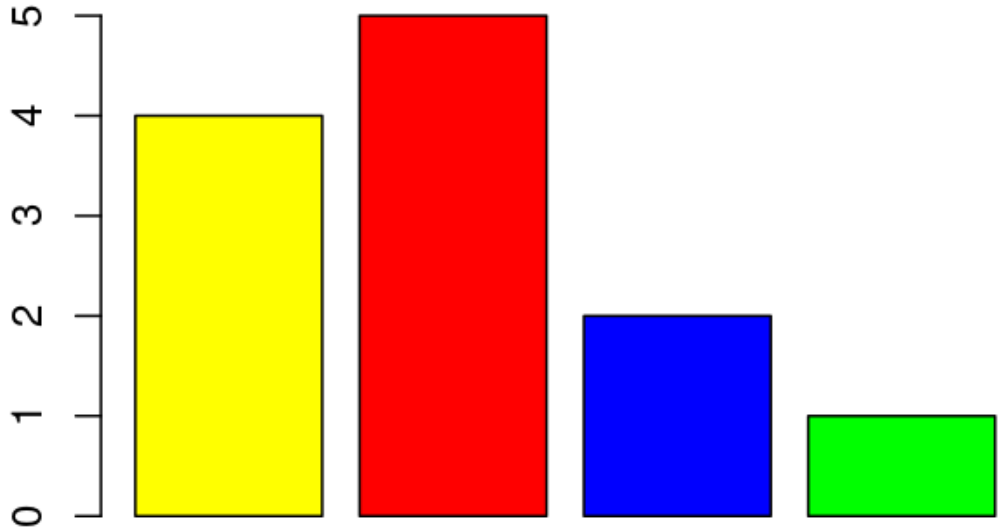
```
#e)(ii) Box Plot 2
boxplot(weight~feed, data=chickwts)
```



F. BAR PLOT

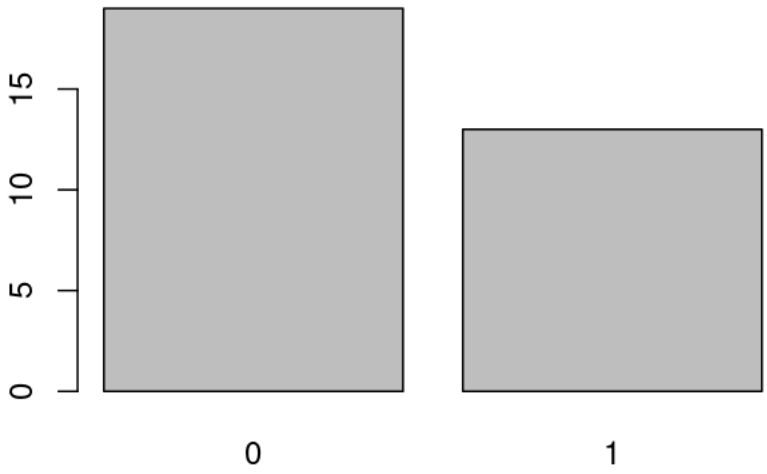
1. BAR PLOT 1

```
#f)(i) Bar Plot 1
a <-c(4,5,2,1)
barplot(a,col=c("yellow","red","blue","green"))
```



2. BAR PLOT 2

```
#f)(ii) Bar Plot 2
cars = mtcars$am
table(mtcars$am)
barplot(table(mtcars$am))
```



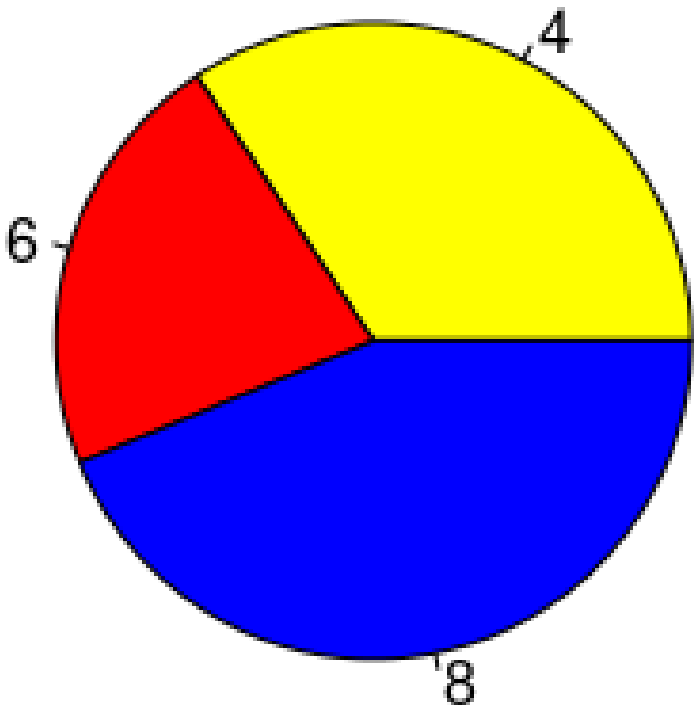
G. PIE CHART

1. PIE CHART 1

```
#g) Pie Chart 1
a <-c(4,5,2,1)
pie(a,col=c("yellow","red","blue","green"))
```

2. PIE CHART 2

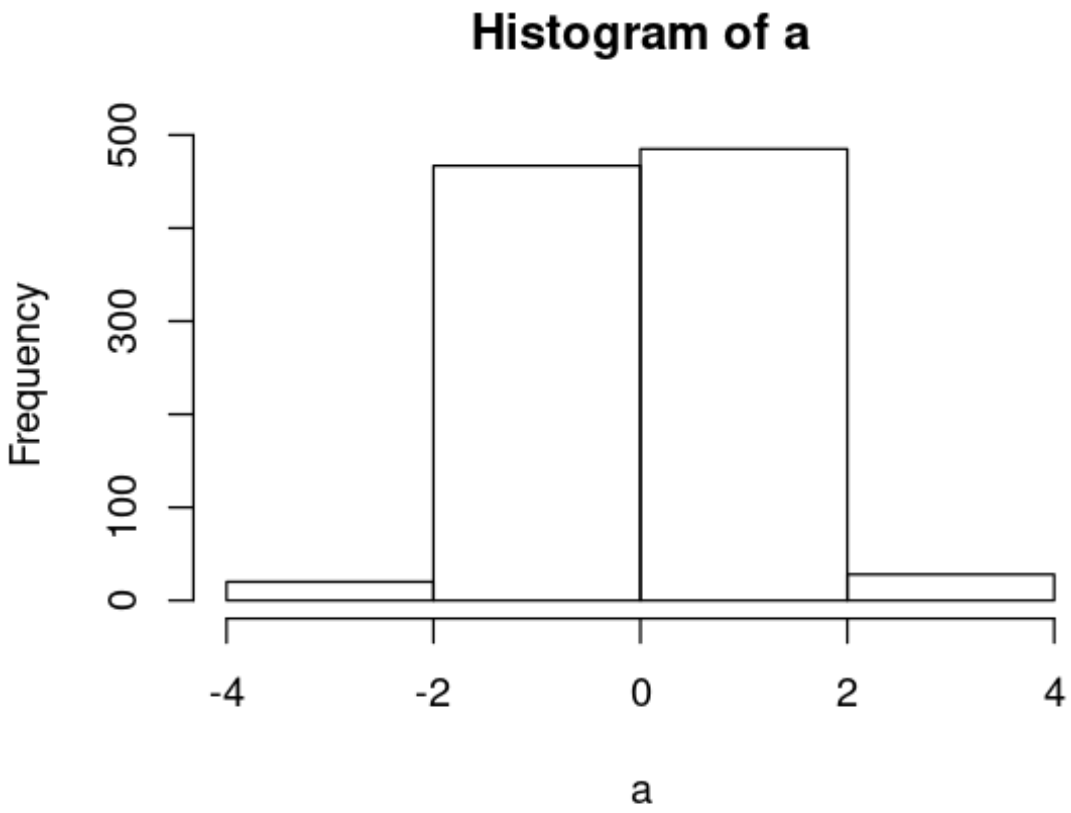
```
#h) Pie Chart 2
cars = table(mtcars$cyl)
pie(cars, col=c("yellow","red","blue"))
```



H. HISTOGRAM

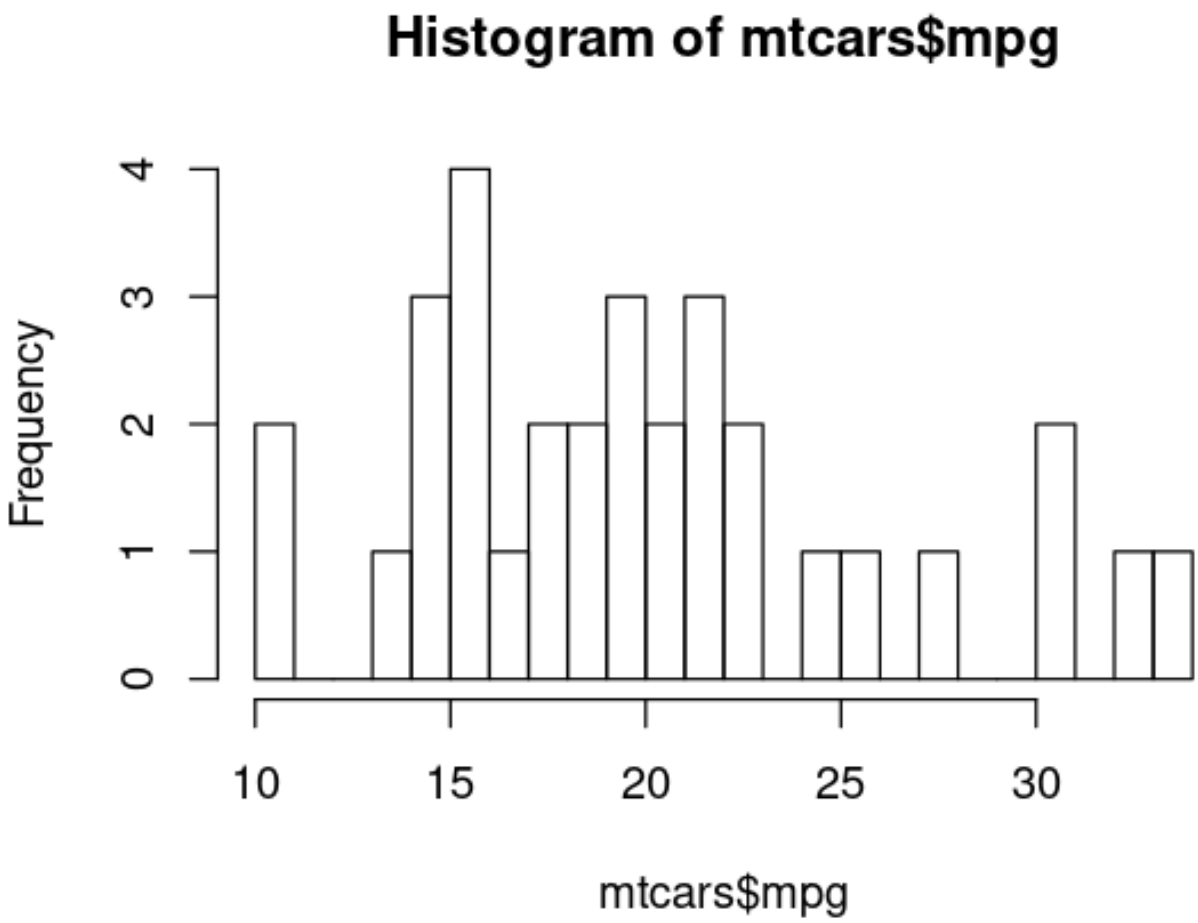
1. HISTOGRAM 1

```
#i) Histogram 1
a <- rnorm(1000)
hist(a,breaks = c(-4,-2,0,2,4))
```



2. HISTOGRAM 2

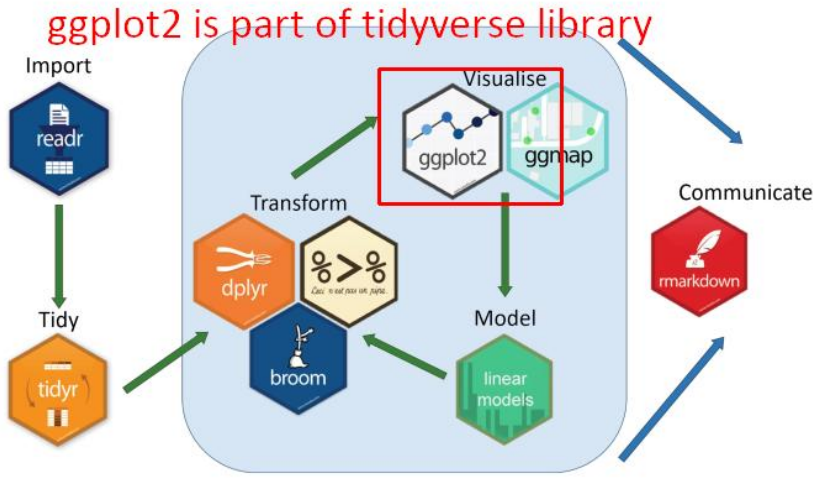
```
#j) Histogram 2
hist(mtcars$mpg, breaks=30)
```



https://www.alvinang.sg/s/Data-Visualisation-with-GGPLOT-R-by-Dr-Alvin-Ang.R

A. INSTALLING TIDYVERSE

- Tidyverse: https://www.tidyverse.org/



- Tidyverse is a R package that contains many libraries: ggplot 2 is one of them.
- You need to install tidyverse package before you can use ggplot 2
- **install.packages("tidyverse", dependencies=TRUE)**

```
#A) Install TIDYVERSE
install.packages("tidyverse", dependencies=TRUE)
```

1. INSTALLIING TIDYVERSE INTO LINUX MINT

- You most probably have no issues installing Tidyverse into R using Windows.
- But Linux Mint is tough.
- Do the following:
    - sudo apt install g++
    - sudo apt-get update
    - sudo apt-get install libcurl4-openssl-dev
    - sudo apt-get install r-base-dev.
    - reboot your laptop
    - reinstall tidyverse:
        - install.packages("tidyverse", dependencies=TRUE)
    - sudo apt install libssl-dev libxml2-dev

B. A) IMPORTING LIBRARIES

Load in these libraries to use ggplot2:

- library(tidyverse)
- library(tibble)
- library(tidyr)
- library(dplyr)
- library(readxl)
- library(ggplot2)
- library(lubridate)

```
#B) Importing Libraries
library(tidyverse)
library(tibble)
library(tidyr)
library(dplyr)
library(readxl)
library(ggplot2)
library(lubridate)
```

C. SCATTER PLOT

1. SCATTER PLOT 1

```
#C1) Scatter Plot 1
# Read the college dataset
# The file can be found here: https://www.alvinang.sg/s/college.csv
college <- read_csv('college.csv')
```

```
# Take a look at the data
summary(college)
```

```
> # Take a look at the data
> summary(college)
       id              name               city              state              region
 Min.   :100654   Length:1269        Length:1269        Length:1269        Length:1269
 1st Qu.:153250   Class :character   Class :character   Class :character   Class :character
 Median :186283   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :186988
 3rd Qu.:215284
 Max.   :484905
 highest_degree       control             gender          admission_rate      sat_avg
 Length:1269        Length:1269        Length:1269        Min.   :0.0509   Min.   : 720
 Class :character   Class :character   Class :character   1st Qu.:0.5339   1st Qu.: 973
 Mode  :character   Mode  :character   Mode  :character   Median :0.6687   Median :1040
                                                          Mean   :0.6501   Mean   :1060
                                                          3rd Qu.:0.7859   3rd Qu.:1120
                                                          Max.   :1.0000   Max.   :1545
   undergrads        tuition      faculty_salary_avg loan_default_rate   median_debt
 Min.   :   47   Min.   : 2732   Min.   : 1451      Length:1269        Min.   : 6056
 1st Qu.: 1296   1st Qu.: 8970   1st Qu.: 6191      Class :character   1st Qu.:21250
 Median : 2556   Median :20000   Median : 7272      Mode  :character   Median :24588
```

notice that all the column summary are useless... they don't furnish much information.... just the total number of rows and its "string character"...

```r
# Convert state, region, highest_degree, control, and gender to factors
college <- college %>%
  mutate(state=as.factor(state),
         region=as.factor(region),
         city=as.factor(city),
         highest_degree=as.factor(highest_degree),
         control=as.factor(control),
         gender=as.factor(gender))
```

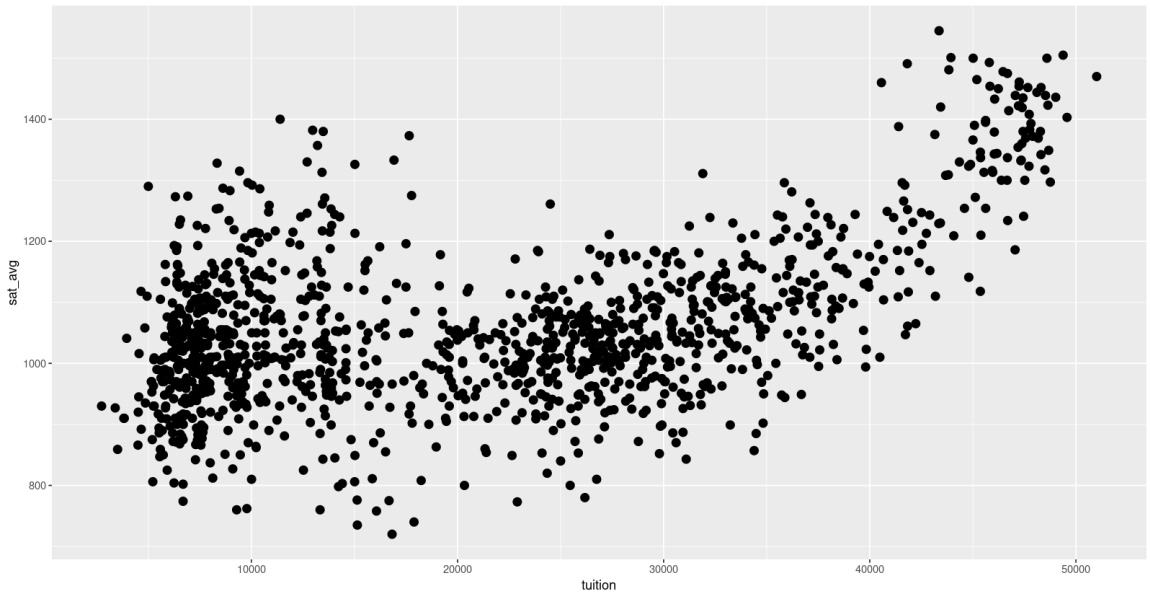
```r
# Take a look at the data
summary(college)
```

```
> # Take a look at the data
> summary(college)
```

its better now... as.factor() changes the column to categories..
and gives a count of them...

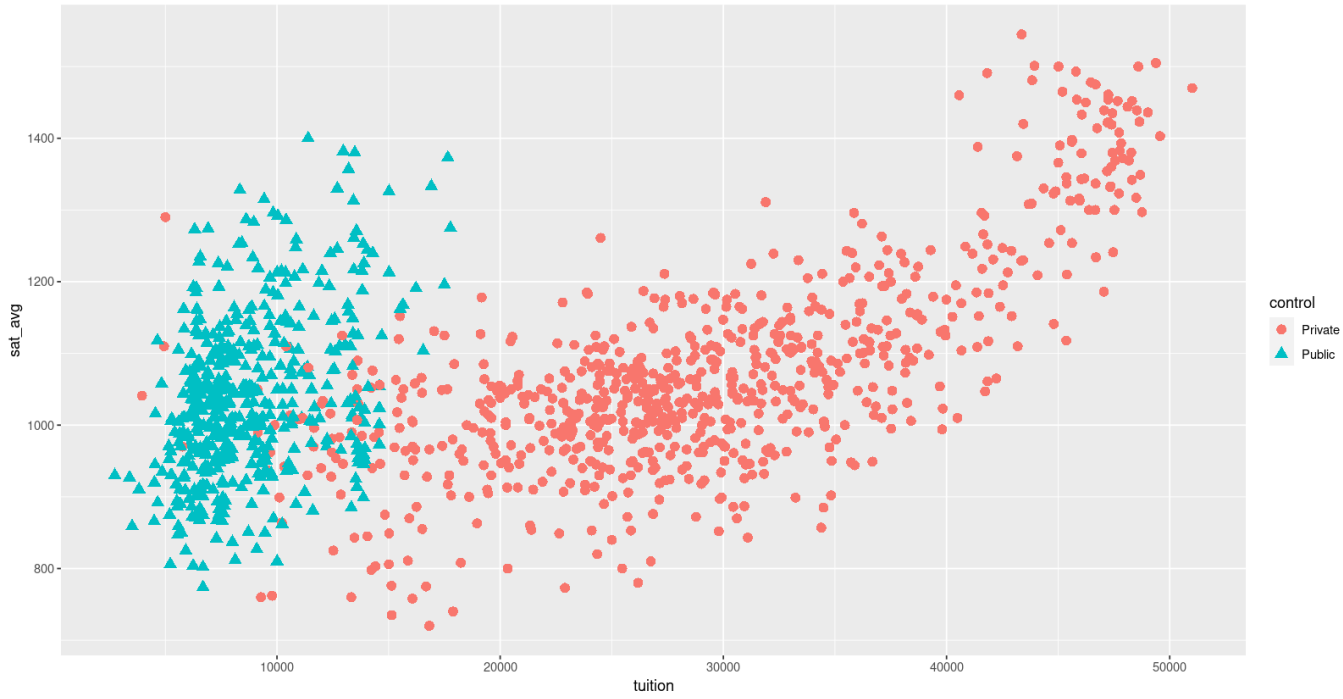
```
       id              name                 city          state          region
 Min.   :100654   Length:1269        New York    :  15   PA     :101   Midwest  :353
 1st Qu.:153250   Class :character   Boston      :  11   NY     : 84   Northeast:299
 Median :186283   Mode  :character   Chicago     :  10   CA     : 71   South    :459
 Mean   :186988                      Philadelphia:   9   TX     : 63   West     :158
 3rd Qu.:215284                      Cleveland   :   8   OH     : 52
 Max.   :484905                      Los Angeles :   8   IL     : 47
                                     (Other)     :1208   (Other):851

   highest_degree    control      gender     admission_rate      sat_avg       undergrads
 Associate:  20   Private:763   CoEd :1237   Min.   :0.0509   Min.   : 720   Min.   :   47
 Bachelor : 200   Public :506   Men  :   4   1st Qu.:0.5339   1st Qu.: 973   1st Qu.: 1296
 Graduate :1049                 Women:  28   Median :0.6687   Median :1040   Median : 2556
                                             Mean   :0.6501   Mean   :1060   Mean   : 5629
                                             3rd Qu.:0.7859   3rd Qu.:1120   3rd Qu.: 6715
                                             Max.   :1.0000   Max.   :1545   Max.   :52280

    tuition      faculty_salary_avg loan_default_rate   median_debt         lon
 Min.   : 2732   Min.   : 1451      Length:1269        Min.   : 6056   Min.   :-157.92
 1st Qu.: 8970   1st Qu.: 6191      Class :character   1st Qu.:21250   1st Qu.: -94.17
 Median :20000   Median : 7272      Mode  :character   Median :24588   Median : -84.89
```

```r
# Let's build a simple scatterplot with tuition on the x-axis
# and average SAT score on the y axis
ggplot(data=college) + geom_point(size=3) + aes(x=tuition,
                                                y=sat_avg)
```



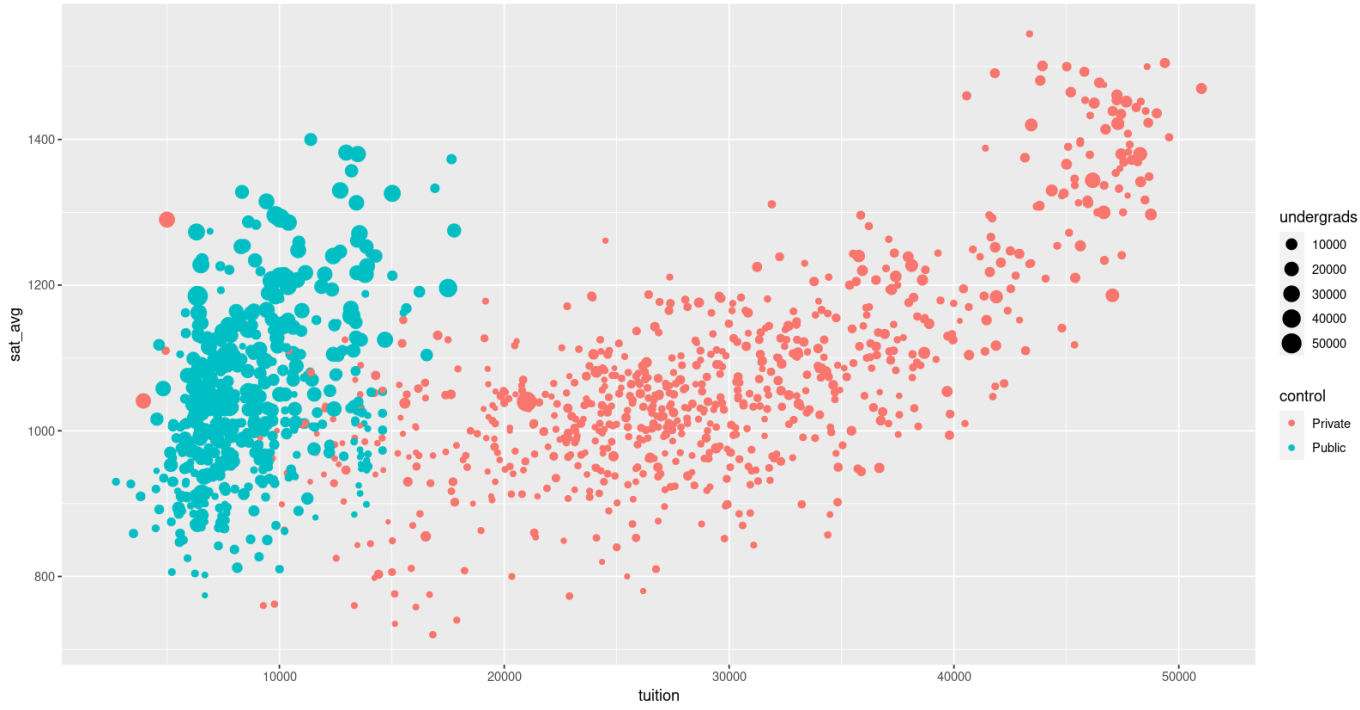
2. SCATTER PLOT 2

```
#C2) Scatter Plot 2
ggplot(data=college) + geom_point(size=3) + aes(x=tuition,
                                                y=sat_avg,
                                                shape=control,
                                                color=control)
```



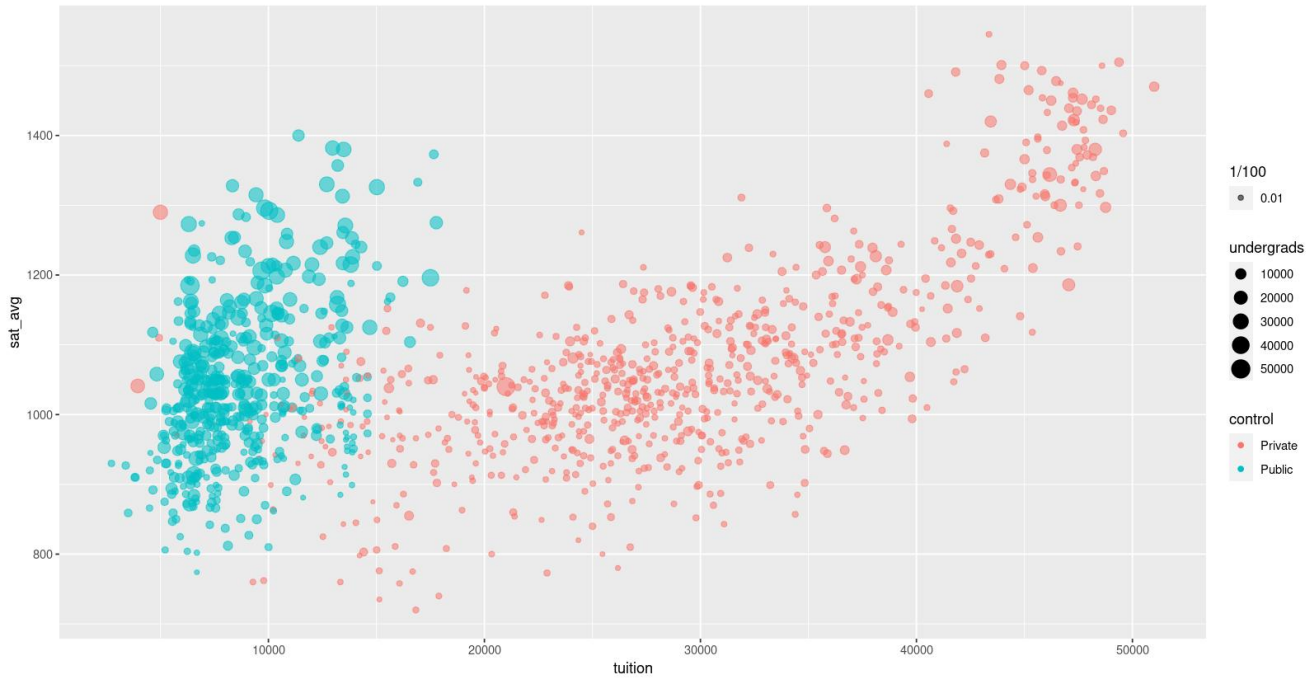
3. SCATTER PLOT 3

```
#C3) Scatter Plot 3
ggplot(data=college) + geom_point() + aes(x=tuition,
                                          y=sat_avg,
                                          color=control,
                                          size=undergrads)
```



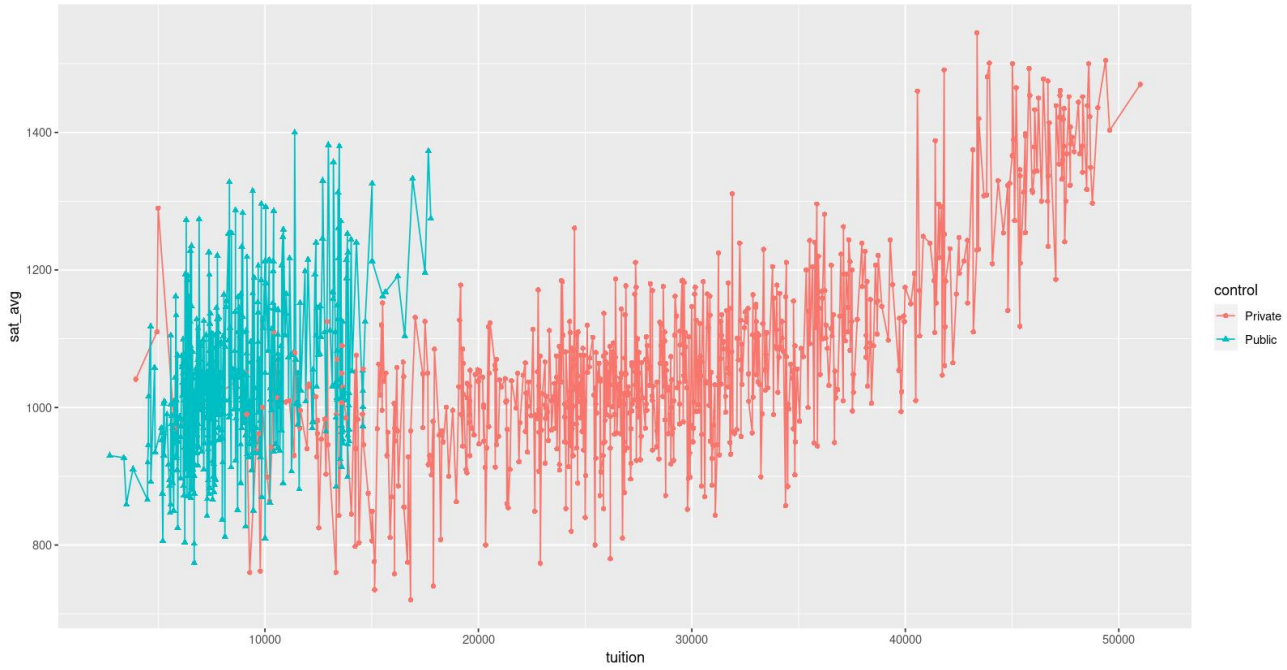
4. SCATTER PLOT 4

```
#C4) Scatter Plot 4
ggplot(data=college) + geom_point() + aes(x=tuition,
                                          y=sat_avg,
                                          color=control,
                                          size=undergrads,
                                          alpha=1/100)
```



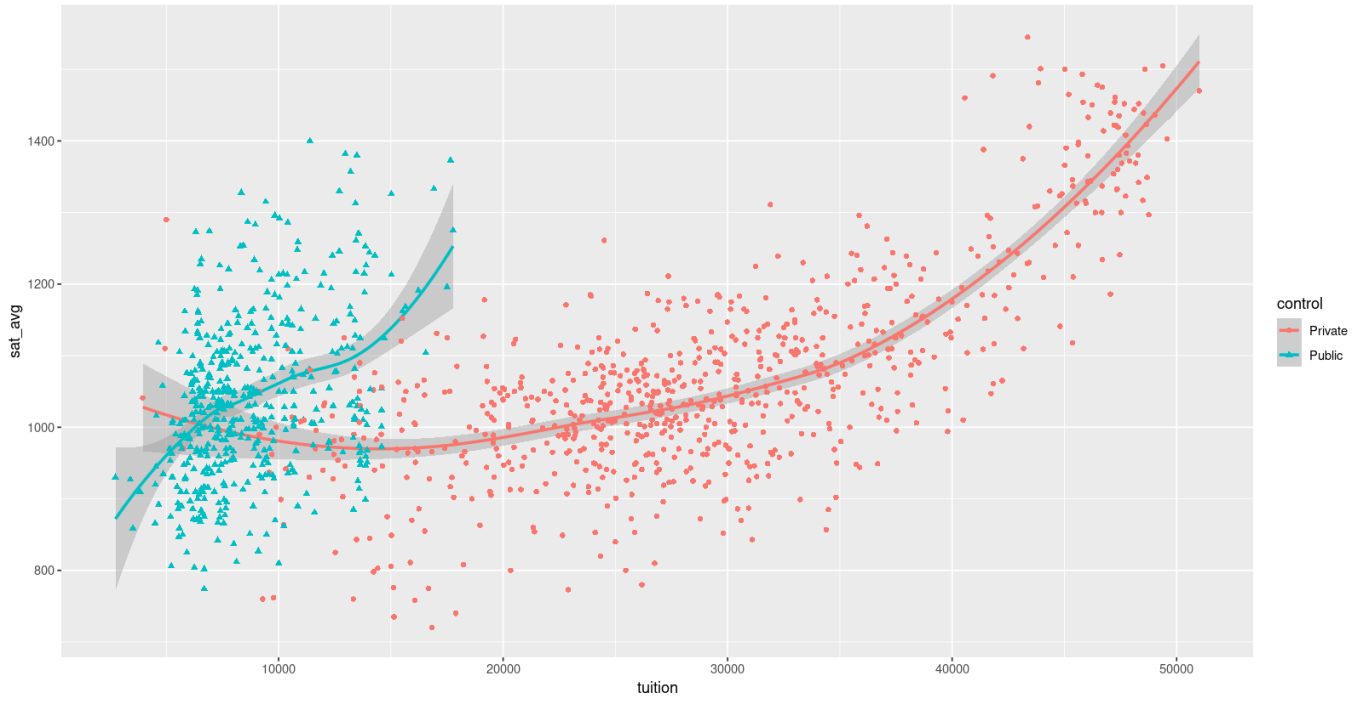
5. ADD LINE TO SCATTER PLOT 1

```
#C5) Add Line to Scatter Plot 1
ggplot(data=college) + geom_line() + geom_point() + aes(x=tuition,
                                                        y=sat_avg,
                                                        shape=control,
                                                        color=control)
```



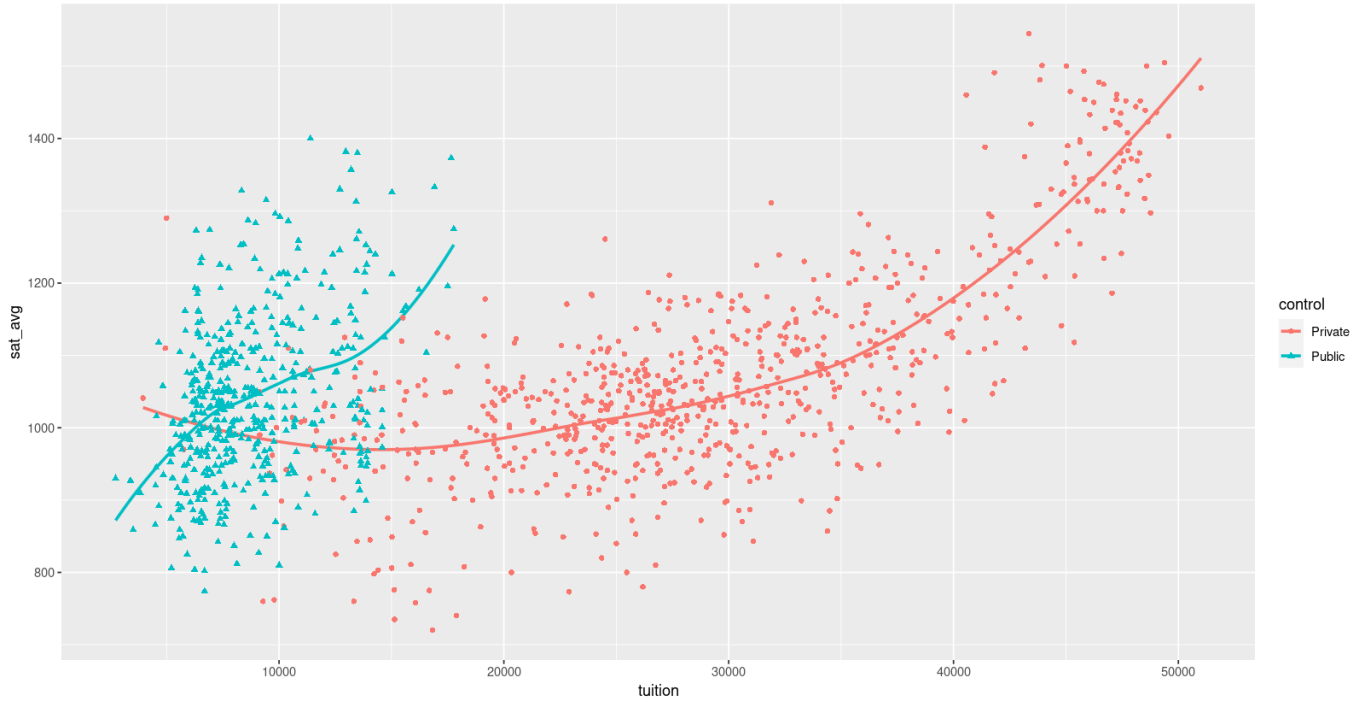
6. ADD LINE TO SCATTER PLOT 2

```
#C6) Add Line to Scatter Plot 2
ggplot(data=college) + geom_smooth() + geom_point() + aes(x=tuition,
                                                          y=sat_avg,
                                                          shape=control,
                                                          color=control)
```



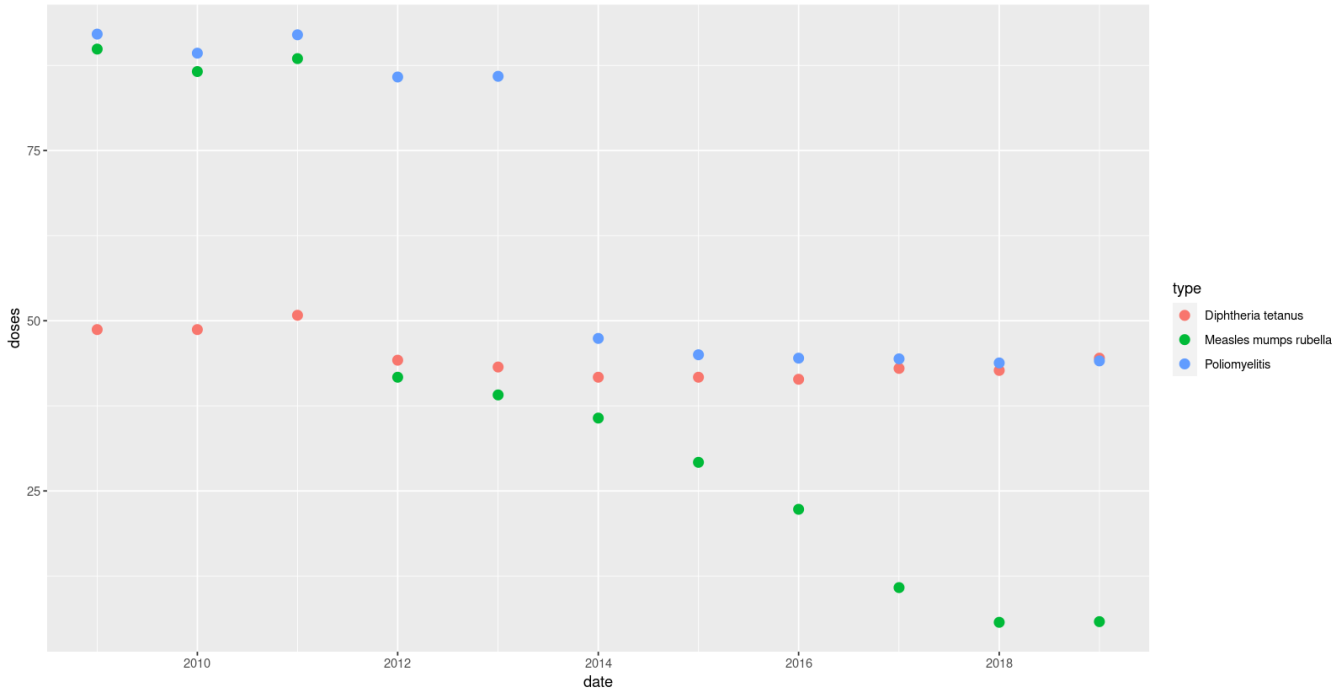
7. ADD LINE TO SCATTER PLOT 3

```
#C7) Add Line to Scatter Plot 3
ggplot(data=college) + geom_smooth(se=FALSE) + geom_point() + aes(x=tuition,
                                                                 y=sat_avg,
                                                                 shape=control,
                                                                 color=control)
```



8. SCATTER PLOT 5

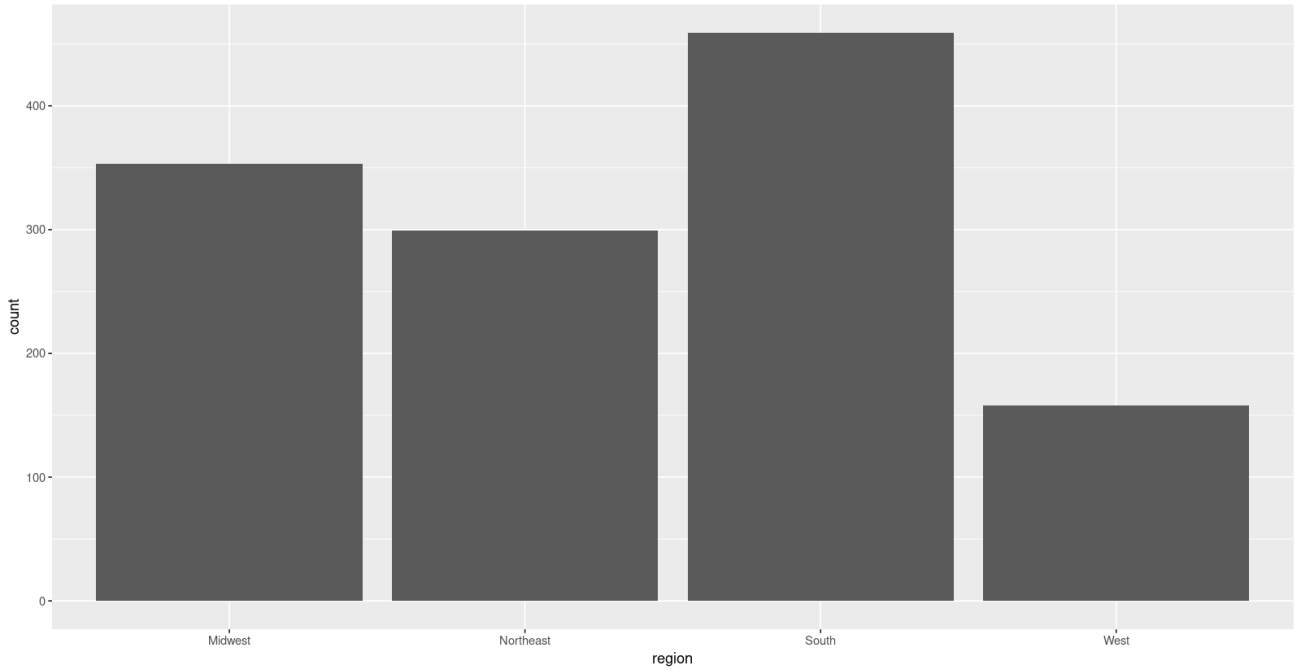
```
#C8) Scatter Plot 5
#File can be found here: https://www.alvinang.sg/s/vaccination.xlsx

vaccination <- read_excel("vaccination.xlsx")

vaccine <- vaccination %>%
  filter(complete.cases(.)) %>%
  mutate(date = ymd(paste0(year,"-01-01"))) %>%
  mutate(doses = no_of_doses_in_thousands) %>%
  mutate(type = vaccination_type) %>%
  select(date,type, doses)

ggplot(data=vaccine) + geom_point(size=3) + aes(x=date, y=doses, color=type)
```



D. BAR CHART

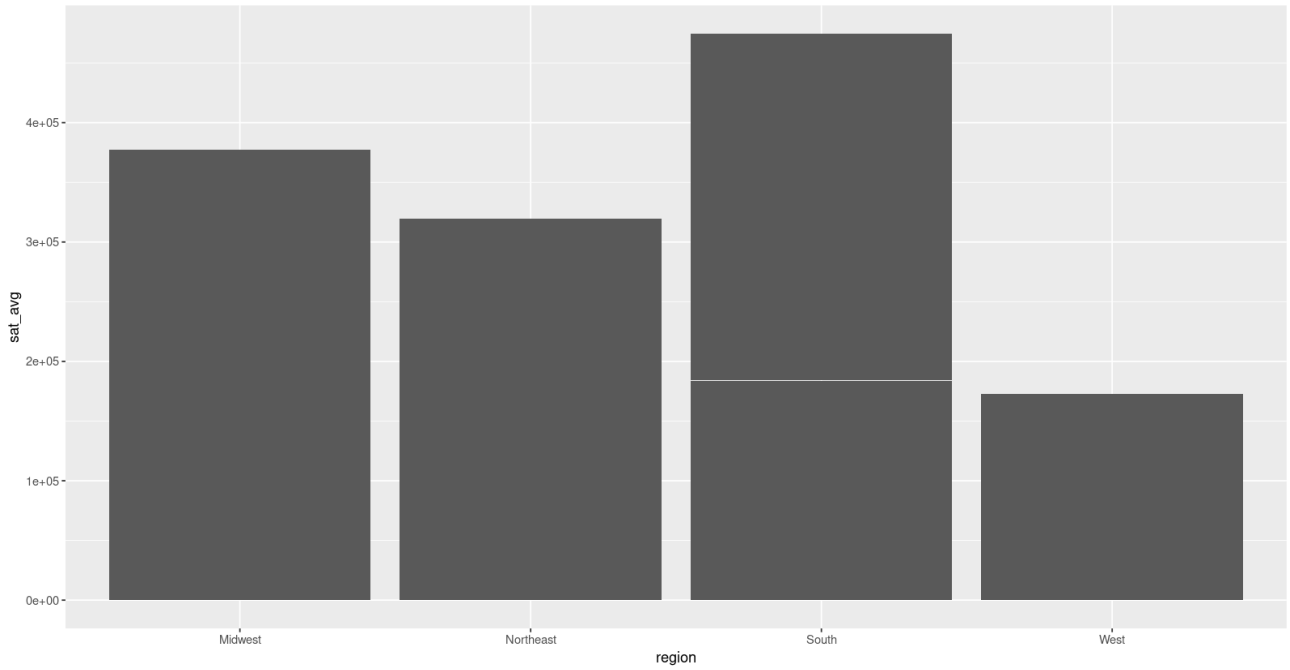
1. BAR CHART 1

```
#D1) Bar Chart 1
ggplot(data=college) + geom_bar() + aes(x=region)
```



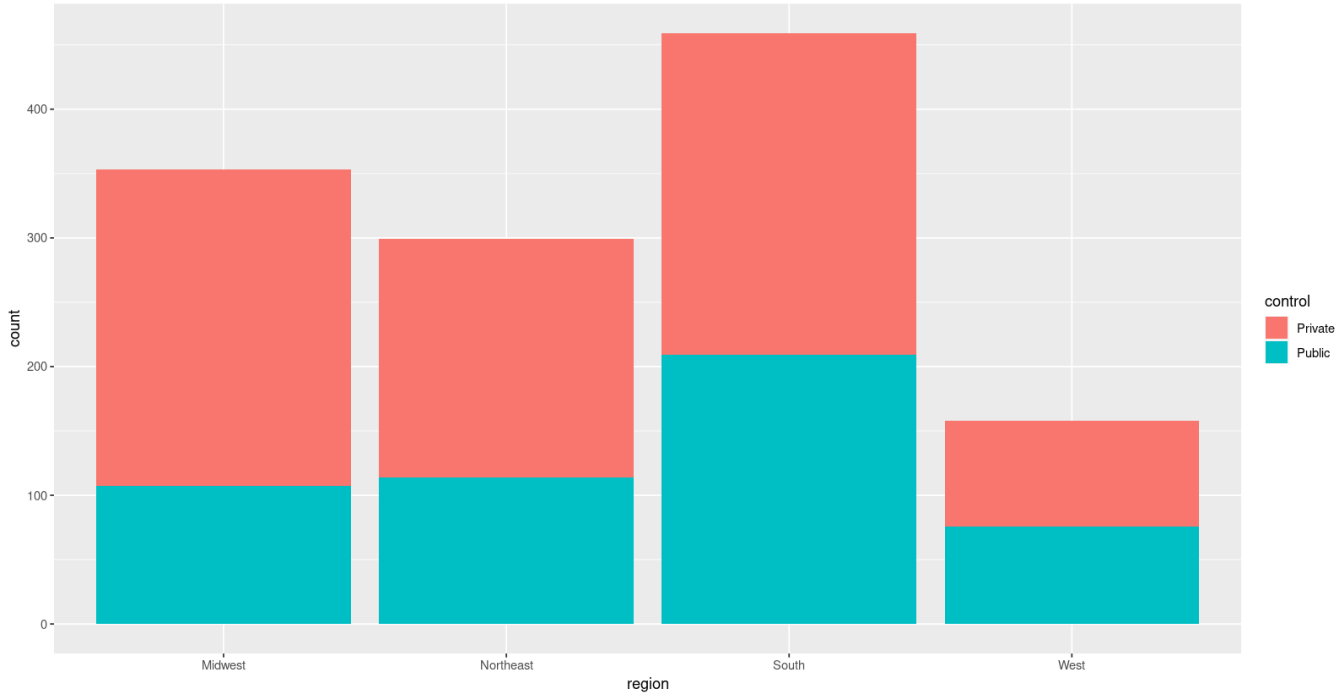
2. BAR CHART 2

```
#D2) Bar Chart 2
ggplot(data=college) + geom_col() + aes(x=region, y = sat_avg)
```



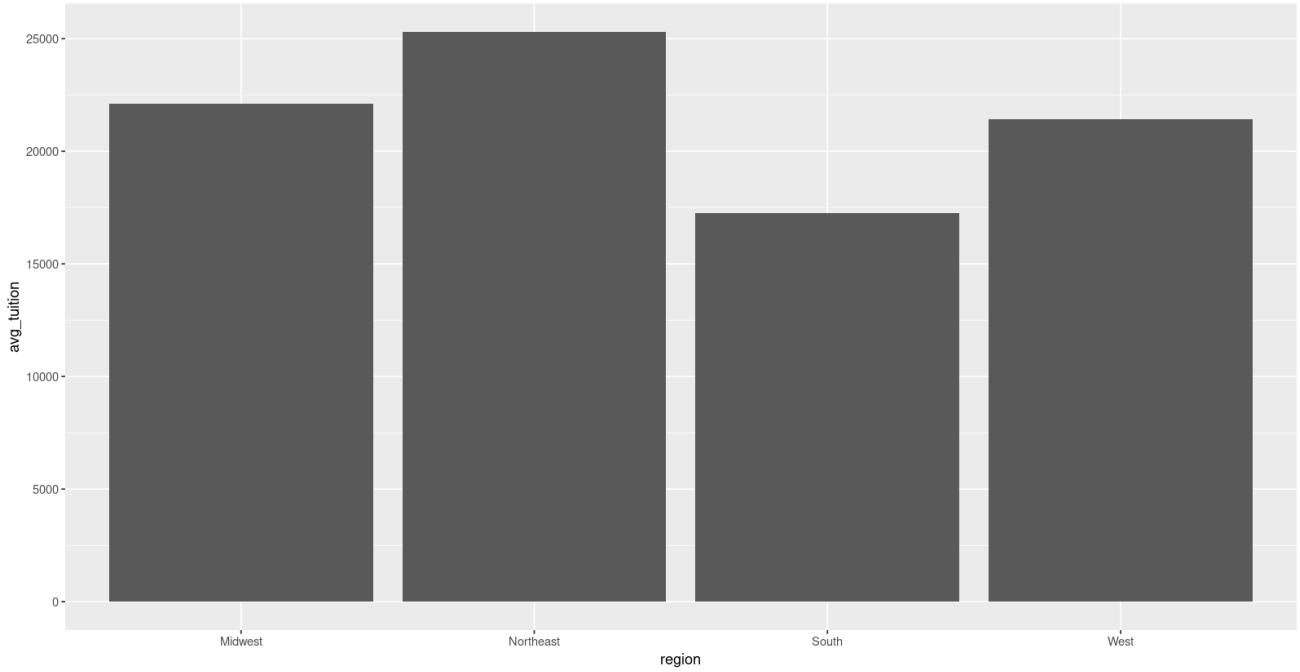
3. **STACKED** BAR CHART 3

```
#D3) Stacked Bar Chart 3
ggplot(data=college) + geom_bar() + aes(x=region, fill=control)
```



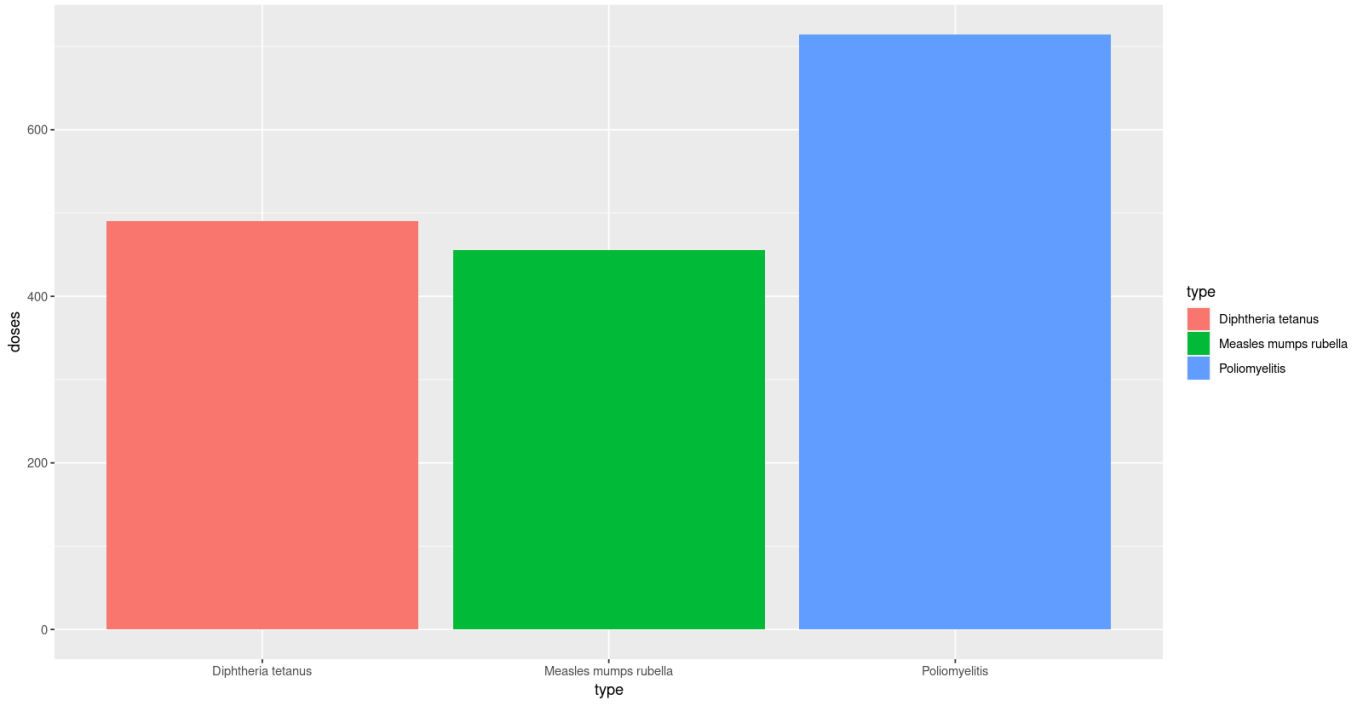
4. BAR CHART 4

```
#D4) Bar Chart 4
college %>%
  group_by(region) %>%
  summarize(avg_tuition=mean(tuition)) %>%
  ggplot() + geom_col() + aes(x=region, y=avg_tuition)
```



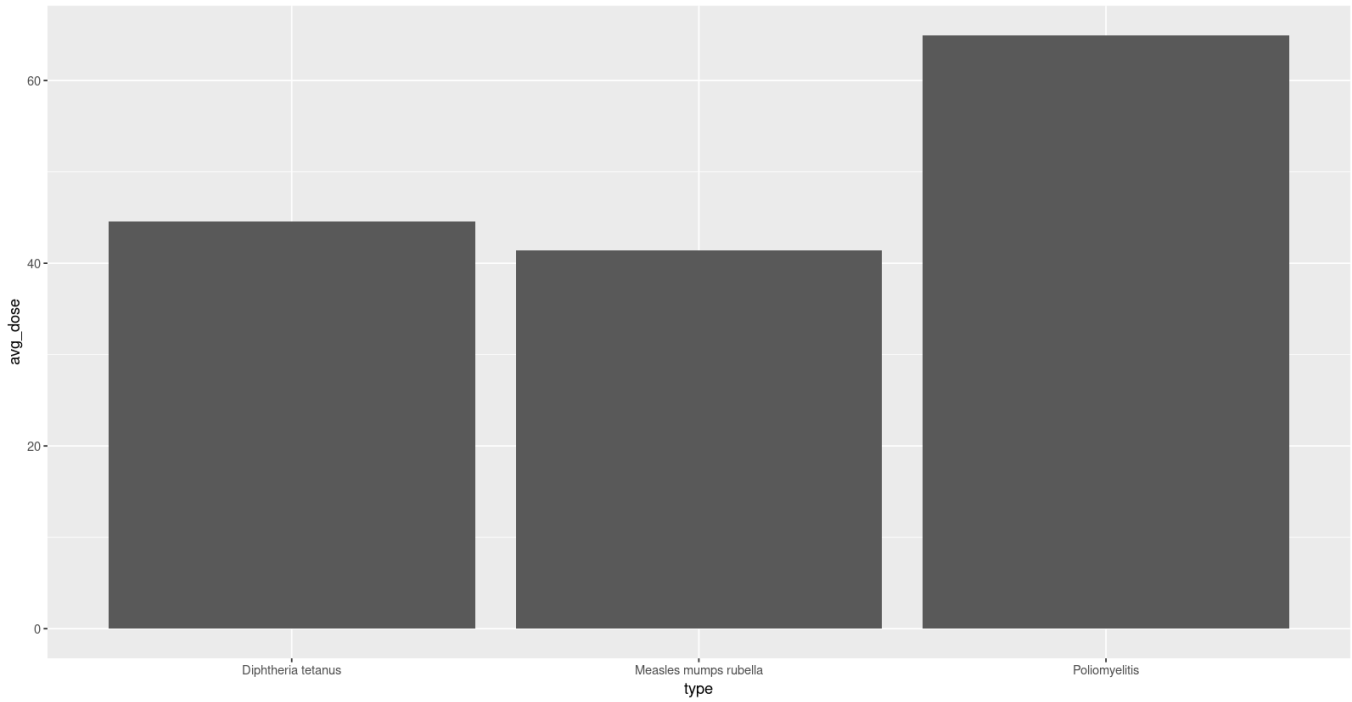
5. BAR CHART 5

```
#D5) Bar Chart 5
ggplot(data=vaccine) + geom_col() + aes(x=type,y=doses,fill=type)
```



6. BAR CHART 6

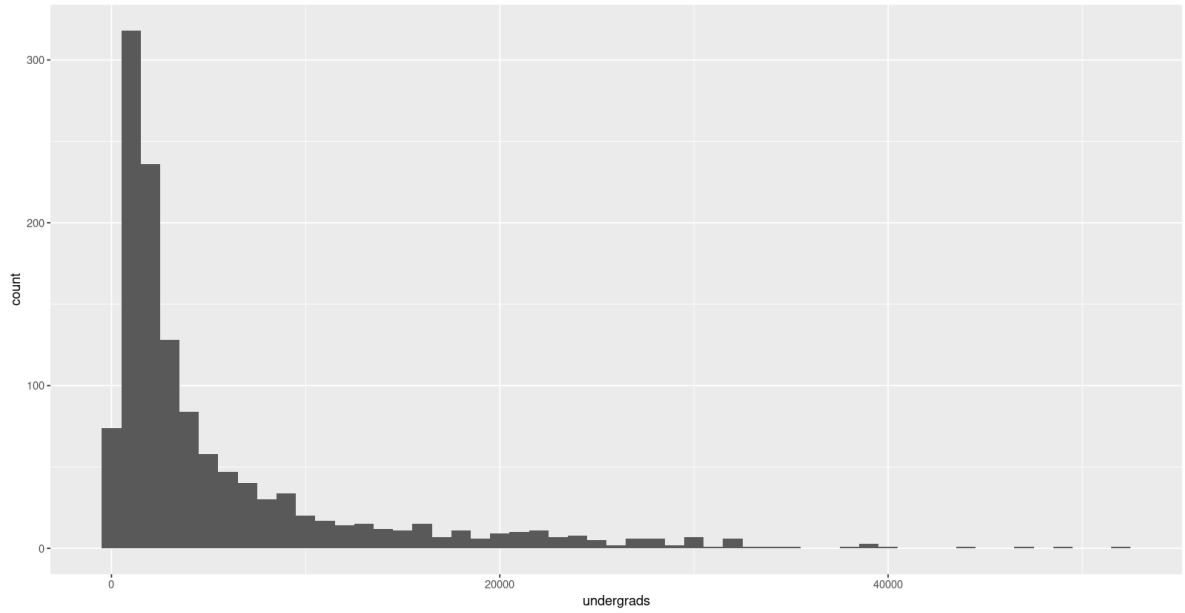
```
#D6) Bar Chart 6
vaccine %>%
  group_by(type) %>%
  summarize(avg_dose=mean(doses)) %>%
  ggplot() + geom_col() + aes(x=type, y=avg_dose)
```



E. HISTOGRAM

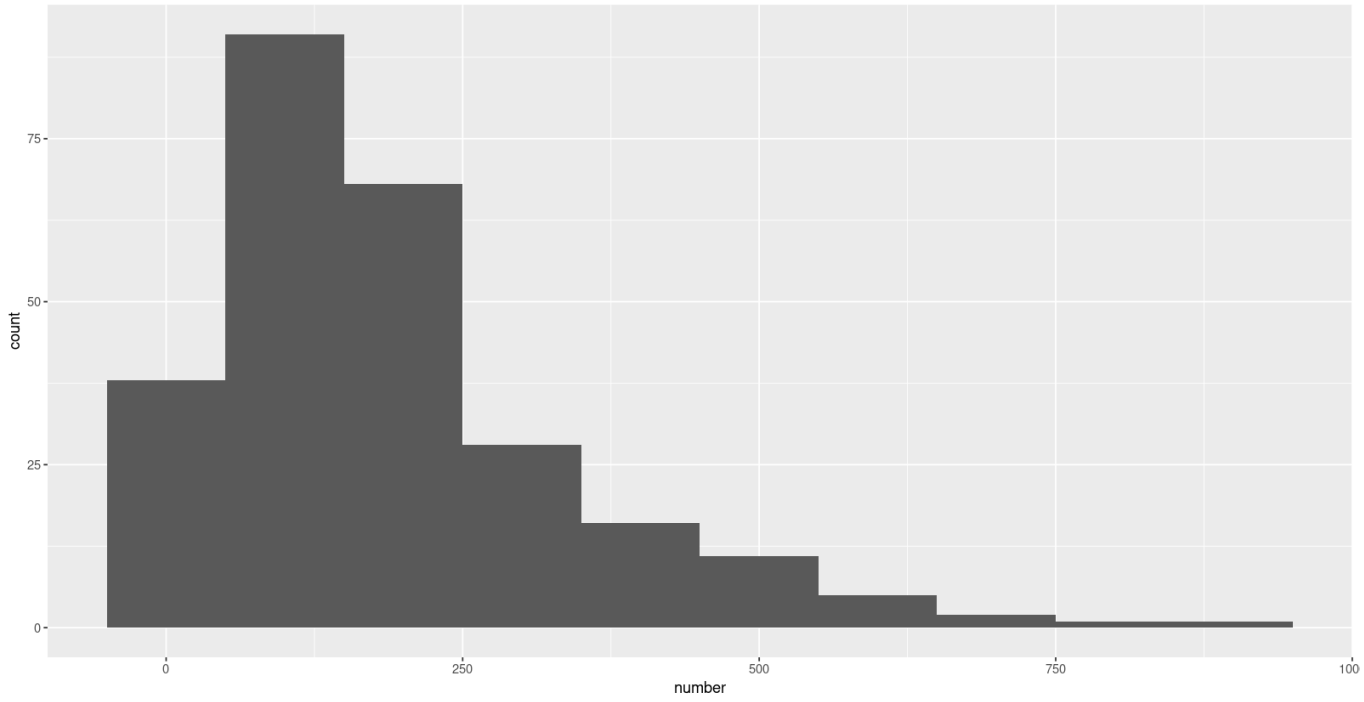
1. HISTOGRAM 1

```
#E1) Histogram 1
ggplot(data=college) + geom_histogram(binwidth=1000) + aes(x=undergrads)
```



2. HISTOGRAM 2

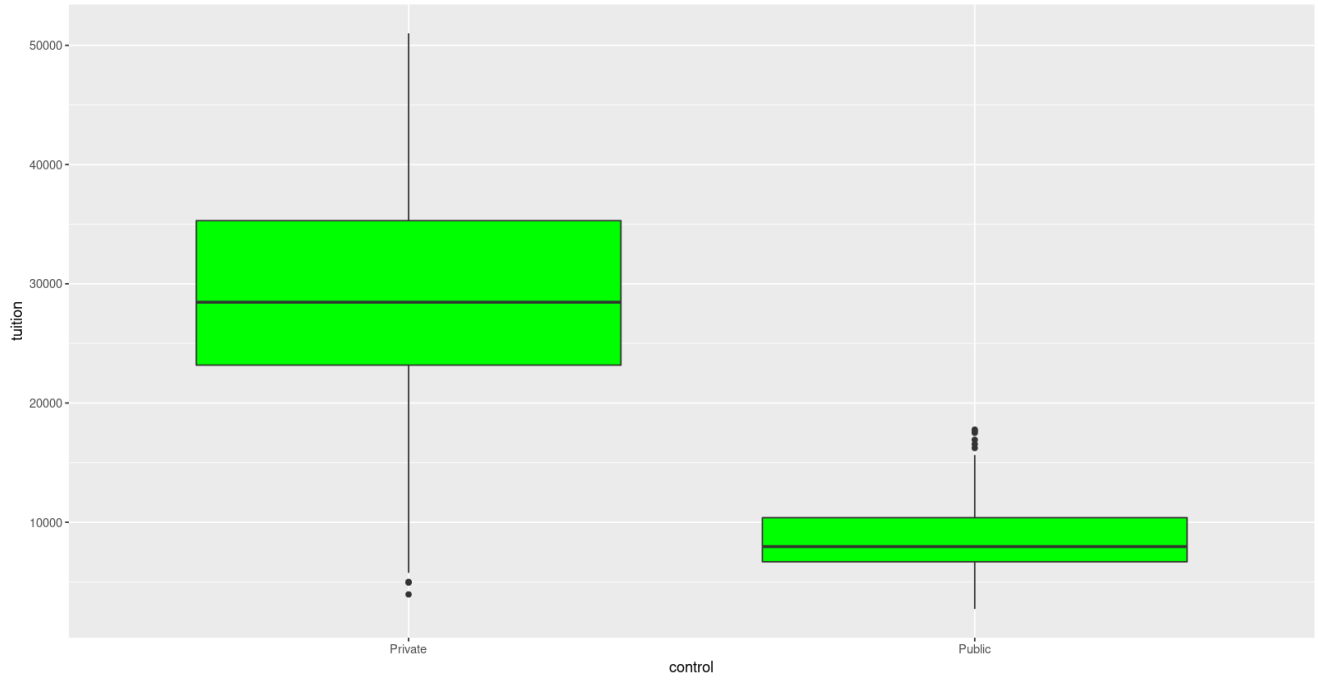
```
#E2) Histogram 2
#File can be found here: https://www.alvinang.sg/s/dengue.csv
dengue <- read_csv("dengue.csv")

dengue %>%
  filter(type_dengue=='Dengue') %>%
  ggplot() + geom_histogram(binwidth =100) + aes(x=number)
```



F. BOX PLOT

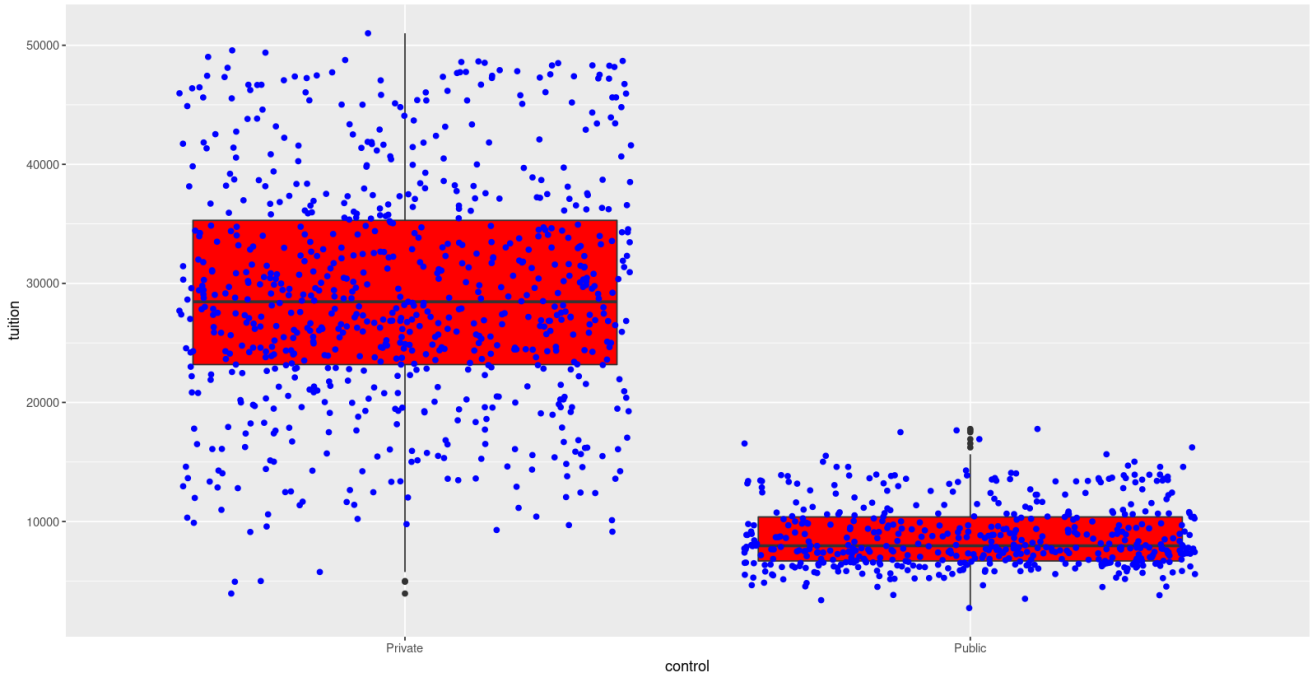
1. BOX PLOT 1

```
#F1) Box Plot 1
ggplot(data=college) + geom_boxplot(fill='green') + aes(x=control, y=tuition)
```



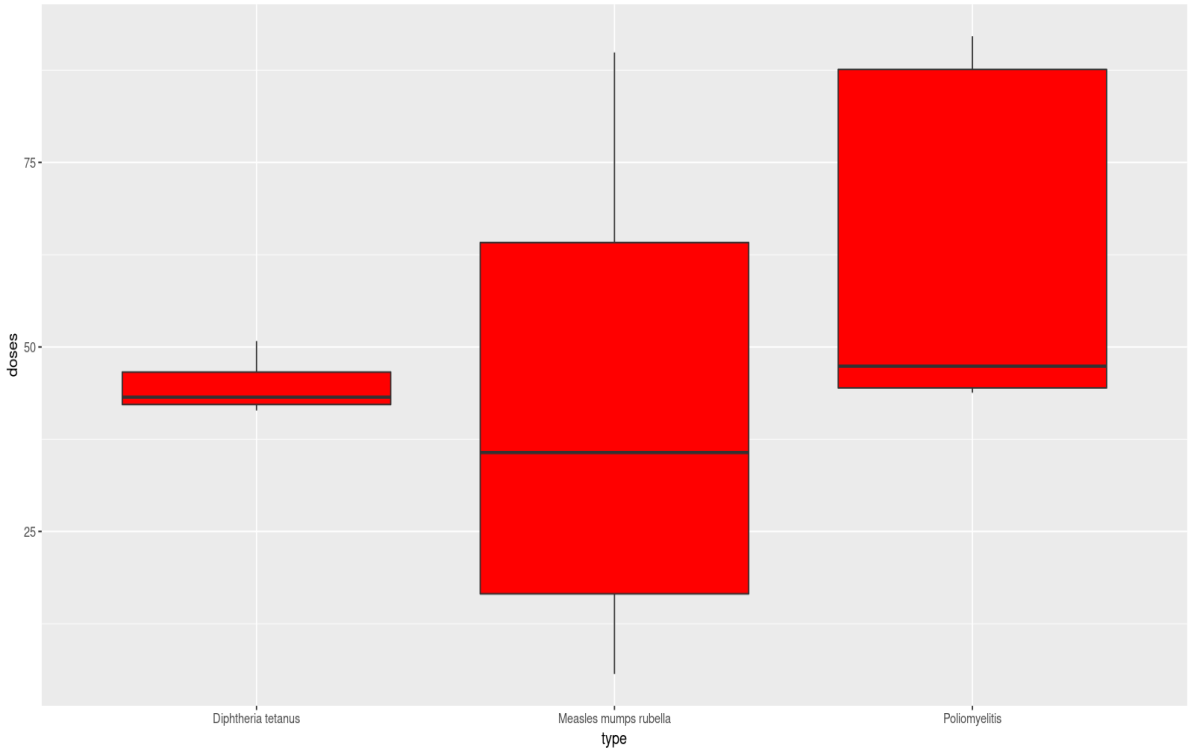
2. BOX PLOT 2

```
#F2) Box Plot 2
ggplot(data=college) + geom_boxplot(fill='red') +
  geom_jitter(col="blue") + aes(x=control, y=tuition)
```



3. BOX PLOT 3

```
#F3) Box Plot 3
vaccination <- read_excel("vaccination.xlsx")

vaccine <- vaccination %>%
  filter(complete.cases(.)) %>%
  mutate(date = ymd(paste0(year,"-01-01"))) %>%
  mutate(doses = no_of_doses_in_thousands) %>%
  mutate(type = vaccination_type) %>%
  select(date,type, doses)

ggplot(data=vaccine) + geom_boxplot(fill='red') + aes(x=type,y=doses)
```



ABOUT DR. ALVIN ANG

Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He is a scientist, entrepreneur, as well as a personal/business advisor. More about him at www.AlvinAng.sg.