Select Split Transform
Data Manipulation

"Practice is the best of all instructors."

PUBLIUS SYRUS, CIRCA 42 B.C

"We all learned by doing, by experimenting (and often failing), and by asking questions."

JAY JACOB WIND

Data Sets

cs2m

grades

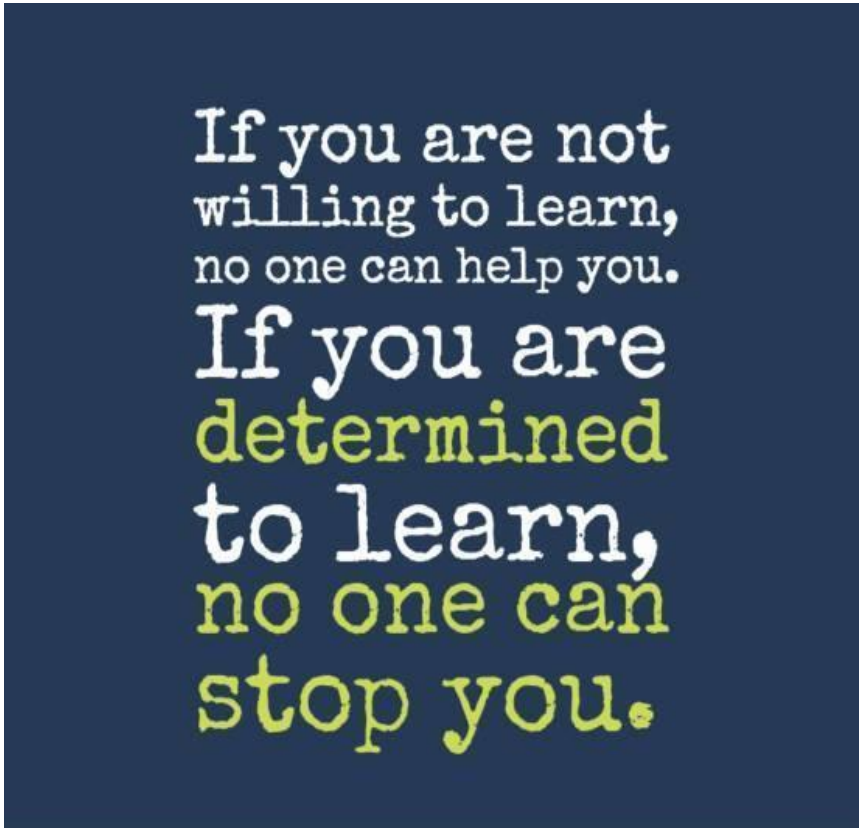
mtcars

preg

t


When You Challenge People,
You Will Lose One Day.
When You Challenge Yourself,
You'll Win Everyday...


You can do it!


If you are not
willing to learn,
no one can help you.
If you are
determined
to learn,
no one can
stop you.

Complete Cases

```
str(t)
summary(t)
```

```
> str(t)
'data.frame':   30 obs. of  6 variables:
 $ BP     : int  100 120 110 100 95 110 120 150 160 125 ...
 $ Chlstrl: int  150 160 150 175 250 200 180 175 185 195 ..
.
 $ Age    : int  20 16 18 25 36 56 59 45 40 20 ...
 $ Prgnt  : int  0 0 0 0 0 0 0 0 0 1 ...
 $ AnxtyLH: int  0 0 0 0 0 NA 1 1 1 0 ...
 $ DrugR  : int  0 0 0 0 0 0 0 0 0 0 ...
> summary(t)
       BP          Chlstrl           Age
 Min.   : 95   Min.   :130.0   Min.   :16.00
 1st Qu.:110   1st Qu.:172.8   1st Qu.:24.00
 Median :125   Median :182.5   Median :33.50
 Mean   :128   Mean   :185.1   Mean   :38.75
 3rd Qu.:145   3rd Qu.:200.0   3rd Qu.:56.00
 Max.   :180   Max.   :250.0   Max.   :81.00
 NA's   :2                     NA's   :2
     Prgnt        AnxtyLH           DrugR
 Min.   :0.0   Min.   :0.0000   Min.   :0.0
 1st Qu.:0.0   1st Qu.:0.0000   1st Qu.:0.0
 Median :0.5   Median :0.0000   Median :0.5
 Mean   :0.5   Mean   :0.4483   Mean   :0.5
 3rd Qu.:1.0   3rd Qu.:1.0000   3rd Qu.:1.0
 Max.   :1.0   Max.   :1.0000   Max.   :1.0
 NA's   :2     NA's   :1
```

Total 7 missing values

complete.cases

```
#-----Complete cases
t_complete = t[complete.cases(t),]
summary(t_complete)
dim(t_complete)
```

```
> summary(t_complete)
       BP            Chlstrl           Age
 Min.   : 95.0   Min.   :130.0   Min.   :16.00
 1st Qu.:115.0   1st Qu.:167.5   1st Qu.:23.00
 Median :125.0   Median :185.0   Median :32.00
 Mean   :129.6   Mean   :186.5   Mean   :37.83
 3rd Qu.:147.5   3rd Qu.:200.0   3rd Qu.:50.50
 Max.   :180.0   Max.   :250.0   Max.   :81.00
     Prgnt           AnxtyLH           DrugR
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :0.4348   Mean   :0.4783   Mean   :0.4783
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

```
> dim(t_complete)
[1] 23  6
```

30-7 = 23

mutate →Creating new variable in data set [run library(dplyr) first]

```
library(dplyr)
cs2m_mutate<- mutate(cs2m, chlst_bp = Chlstrl/BP)
head(cs2m_mutate)
```

```
> cs2m_mutate<- mutate(cs2m, chlst_bp = Chlstrl/BP)
> head(cs2m_mutate)
   BP Chlstrl Age Prgnt AnxtyLH DrugR chlst_bp
1 100     150  20     0       0     0 1.500000
2 120     160  16     0       0     0 1.333333
3 110     150  18     0       0     0 1.363636
4 100     175  25     0       0     0 1.750000
5  95     250  36     0       0     0 2.631579
6 110     200  56     0       1     0 1.818182
```

If you write cs2m, left side, a new variable will be created in the original data set

New df

Create new variable by another method

New data set, a copy of cs2m

New data set, a copy of cs2m

Name of new variable after $ sign

```
> cs2m_1$chlst_bp<- cs2m_1$Chlstrl/cs2m_1$BP
> View(cs2m_1)
```

Values of new variable

217.R × | mpg × | cs2m_1 × | grades × | cs2m × | BasicsR.R ×

Filter

	BP	Chlstrl	Age	Prgnt	AnxtyLH	DrugR	chlst_bp
1	100	175	25	0	0	0	1.750000
2	165	200	25	1	0	0	1.212121
3	145	175	30	1	0	0	1.206897
4	120	180	28	1	0	0	1.500000
5	100	180	21	1	0	0	1.800000
6	120	200	30	1	0	1	1.666667
7	125	240	29	1	0	1	1.920000
8	130	172	30	1	0	1	1.323077

Showing 1 to 8 of 8 entries

Change the name of variables *DrugR* to Reaction and *Prgnt* to Pregnant

```
j = m
dim(j)
install.packages("reshape")
library(reshape)

j = rename(j, c(DrugR = 'Reaction',
                Prgnt = 'Pregnant'))
```

j & m are nothing but cs2m file. First write m = cs2m

```
variable.names(j)
names(j)
```

```
> variable.names(j)
[1] "BP"        "Chlstrl"   "Age"       "Pregnant" "AnxtyLH"
[6] "Reaction"
> names(j)
[1] "BP"        "Chlstrl"   "Age"       "Pregnant" "AnxtyLH"
[6] "Reaction"
```

Using names()

AnxtyLH to Anxiety

Chlstrl to Cholesterol

```
names(j)[5] = "Anxiety"
names(j)[2] = "Cholesterol"

variable.names(j)
```

```
> names(j)
[1] "BP"          "Cholesterol" "Age"
[4] "Pregnant"    "Anxiety"     "Reaction"
```

arrange

By default ascending order (low to high)

```
> cs2m_asce<- arrange(cs2m, Age)
> head(cs2m_asce)
# A tibble: 6 x 6
     BP Chlstrl   Age Prgnt AnxtyLH DrugR
  <int>   <int> <int> <int>   <int> <int>
1   120     160    16     0       0     0
2   110     150    18     0       0     0
3   135     190    18     1       0     0
4    95     250    18     1       0     1
5   100     160    19     1       0     1
6   100     150    20     0       0     0
```

For Descending order (High to Low), need to specify

```
> cs2m_desc<- arrange(cs2m, desc(Age))
> head(cs2m_desc)
# A tibble: 6 x 6
     BP Chlstrl   Age Prgnt AnxtyLH DrugR
  <int>   <int> <int> <int>   <int> <int>
1   180     200    81     0       1     1
2   140     190    73     0       1     1
3   130     175    72     0       1     1
4   150     195    65     0       1     1
5   120     180    59     0       1     0
6   145     210    58     0       1     1
```

Select only quiz1, gpa & final and view few top rows

```
> grades1<-subset(grades, select = c(quiz1, gpa, final))
> head(grades1)
```

	quiz1	gpa	final
1	6	1.18	53
2	10	2.19	54
3	10	2.46	57
4	7	3.98	68
5	7	1.84	66
6	10	3.90	74

select = c(…
c is for concatenate

Much easier way!

```
> grades4<- select(grades, quiz1, gpa, final)
> grades4
# A tibble: 105 x 3
   quiz1   gpa final
   <int> <dbl> <int>
 1     6  1.18    53
 2    10  2.19    54
 3    10  2.46    57
 4     7  3.98    68
 5     7  1.84    66
 6    10  3.90    74
 7    10  2.84    63
 8    10  3.57    71
 9    10  3.95    74
10    10  3.49    75
# ... with 95 more rows
```

apply → column **MEANS**

2 stands for columns

```
> apply(cs2m,2,mean)
          BP     Chlstrl         Age       Prgnt
127.3333333 185.0666667  37.7666667   0.5000000
    AnxtyLH       DrugR
  0.4666667   0.5000000
```

```
> mean(cs2m)
[1] NA
Warning message:
In mean.default(cs2m) : argument is not numeric
 or logical: returning NA
> mean(cs2m$BP)
[1] 127.3333
```

apply → row **MEANS**

1 stands for rows

```
> apply(cs2m,1,mean)
 [1] 45.00000 49.33333 46.33333 50.00000
 [5] 63.50000 61.16667 60.00000 61.83333
 [9] 64.33333 56.83333 57.33333 65.16667
[13] 58.50000 54.83333 50.33333 46.83333
[17] 60.83333 58.66667 66.00000 55.66667
[21] 48.00000 50.16667 53.33333 57.16667
[25] 68.66667 63.16667 71.33333 69.16667
[29] 77.16667 67.50000
```

This makes no sense

Average of *all columns* by ***cylinder***

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
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
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2

See what is at 2nd number in data, cyl, you donot want this to be averaged

4th cylinder

```
> by(mtcars[,-2], mtcars$cyl, colMeans)
mtcars$cyl: 4
        mpg        disp          hp
 26.6636364 105.1363636  82.6363636
       drat          wt        qsec
  4.0709091   2.2857273  19.1372727
         vs          am        gear
  0.9090909   0.7272727   4.0909091
       carb
  1.5454545
--------------------------------
```

Average of *all columns* by ***cylinder***

```
-----------------------------------
mtcars$cyl: 6
        mpg        disp          hp
 19.7428571 183.3142857 122.2857143
       drat          wt        qsec
  3.5857143   3.1171429  17.9771429
         vs          am        gear
  0.5714286   0.4285714   3.8571429
       carb
  3.4285714
-----------------------------------
```

6th cylinder

8th cylinder

```
-----------------------------------
mtcars$cyl: 8
        mpg        disp          hp
 15.1000000 353.1000000 209.2142857
       drat          wt        qsec
  3.2292857   3.9992143  16.7721429
         vs          am        gear
  0.0000000   0.1428571   3.2857143
       carb
  3.5000000
```

One Variable's mean across a categorical variable

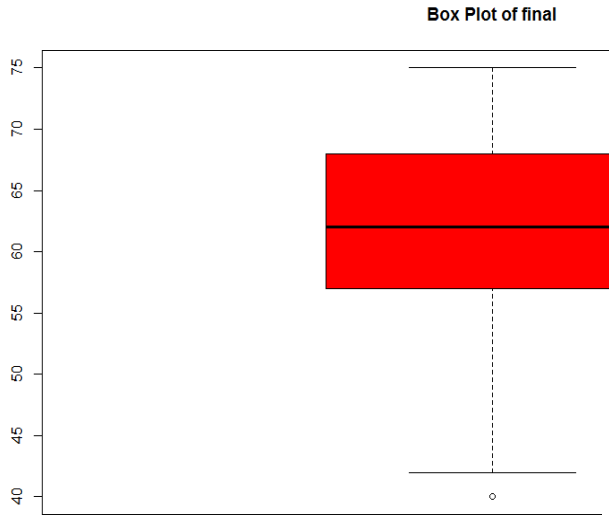
```
> tapply(cs2m$BP, cs2m$Prgnt, mean)
       0        1
132.0000 122.6667
> tapply(grades$gpa, grades$ethnicity, mean)
       1        2        3        4        5
2.966000 2.557000 2.698333 2.872889 2.888182
```

Select only final > 60 and view few top rows

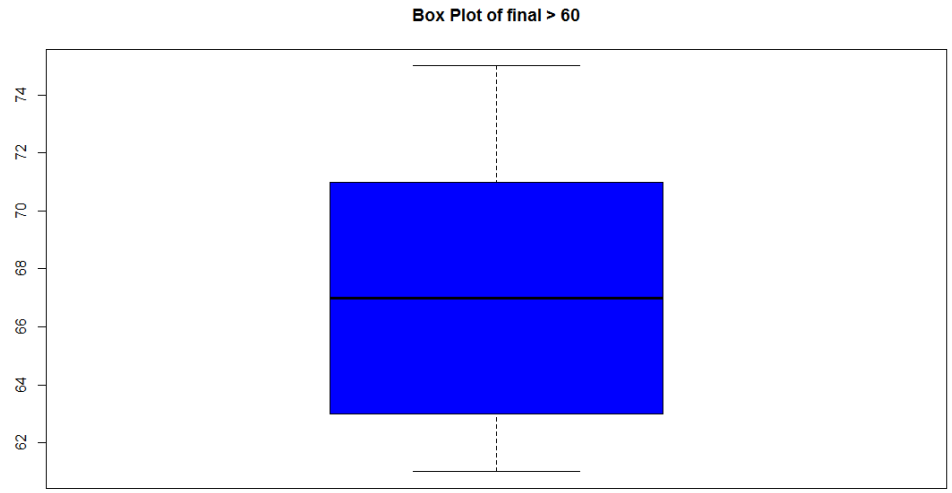
```
> final_60<- subset(grades, final>60)
> head(final_60)
  Sr_No     id lastname firstname gender ethnicity year lowup section  gpa
4     4 132931  OSBORNE       ANN      1         3    2     1       2 3.98
5     5 140219   GUADIZ   VALERIE      1         2    4     2       1 1.84
6     6 142630  RANGIFO   TANIECE      1         4    3     2       3 3.90
7     7 153964 TOMOSAWA    DANIEL      2         2    3     2       3 2.84
8     8 154441     LIAN     JENNY      1         5    2     1       1 3.57
9     9 157147   BAKKEN      KREG      2         4    3     2       1 3.95
  extrc review quiz1 quiz2 quiz3 quiz4 quiz5 final total percent grade
4     1      1     7     8     7     7     6    68   103      82     B
5     1      1     7     8     9     8    10    66   108      86     B
6     1      2    10    10    10     9     9    74   122      98     A
7     2      1    10     9    10    10    10    63   112      90     A
8     1      2    10     9    10    10    10    71   120      96     A
9     2      2    10    10    10    10     9    74   123      98     A
  passfail
4        P
5        P
6        P
7        P
8        P
9        P
```

Mathematical argument

Compare box plots of final of all 105 observations and with final>60




```
> boxplot(grades$final,
main = "Box Plot of final",
col="red")
```



Compare correlation between *final* and gpa in all 105 observations and in subset *final > 60*

```
> cor.test(grades$gpa, grades$final)

        Pearson's product-moment correlation

data:  grades$gpa and grades$final
t = 5.8291, df = 103, p-value = 6.44e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3387243 0.6296171
sample estimates:
      cor
0.498055
```

```
> cor.test(final_60$gpa, final_60$final)

        Pearson's product-moment correlation

data:  final_60$gpa and final_60$final
t = 5.1973, df = 58, p-value = 2.738e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3615041 0.7152358
sample estimates:
      cor
0.5636854
```

cs2m file; **Age** between 20 & 32

filter

```
cs2m_1<- filter(cs2m, Age>20 & Age<32)
cs2m_1
```

```
> cs2m_1
   BP Chlstrl Age Prgnt AnxtyLH DrugR
1 100     175  25     0       0     0
2 165     200  25     1       0     0
3 145     175  30     1       0     0
4 120     180  28     1       0     0
5 100     180  21     1       0     0
6 120     200  30     1       0     1
7 125     240  29     1       0     1
8 130     172  30     1       0     1
```

cs2m file; **Age** between 20 & 32

subset

```
cs2m_2<- filter(cs2m, Age>20 & Age<32)
cs2m_2
```

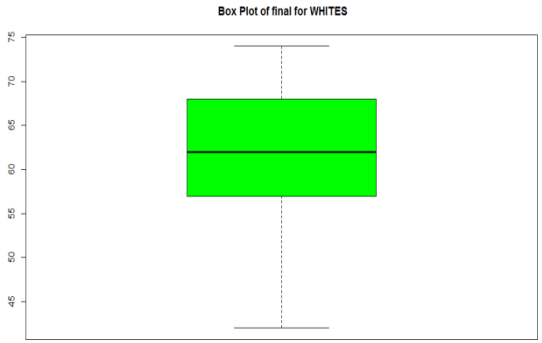
```
> cs2m_2
   BP Chlstrl Age Prgnt AnxtyLH DrugR
1 100     175  25     0       0     0
2 165     200  25     1       0     0
3 145     175  30     1       0     0
4 120     180  28     1       0     0
5 100     180  21     1       0     0
6 120     200  30     1       0     1
7 125     240  29     1       0     1
8 130     172  30     1       0     1
```

You can use subset with small change!

```
> cs2m_3<- subset(cs2m, Age > 19 & Age < 31)
> cs2m_3
# A tibble: 10 x 6
      BP Chlstrl   Age Prgnt AnxtyLH DrugR
   <int>   <int> <int> <int>   <int> <int>
 1   100     150    20     0       0     0
 2   100     175    25     0       0     0
 3   125     195    20     1       0     0
 4   165     200    25     1       0     0
 5   145     175    30     1       0     0
 6   120     180    28     1       0     0
 7   100     180    21     1       0     0
 8   120     200    30     1       0     1
 9   125     240    29     1       0     1
10   130     172    30     1       0     1
```

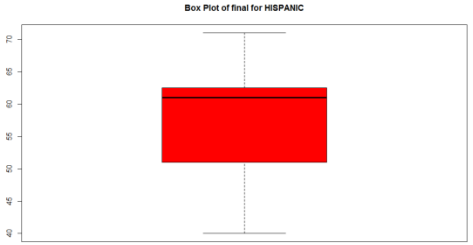
Create subset of only WHITES (ethnicity = 4) and view few top rows and make box plot of final

```
> ethnicity_white<-subset(grades, ethnicity == 4)
> head(ethnicity_white)
   Sr_No     id  lastname firstname gender ethnicity year lowup section  gpa
2      2 108642 VALAZQUEZ     SCOTT      2         4    3     2       2 2.19
3      3 127285    GALVEZ    JACKIE      1         4    4     2       2 2.46
6      6 142630   RANGIFO   TANIECE      1         4    3     2       3 3.90
9      9 157147    BAKKEN      KREG      2         4    3     2       1 3.95
12    12 167664     SWARM      MARK      2         4    3     2       3 2.35
13    13 175325    KHOURY    DENNIS      2         4    3     2       1 2.45
   extrc review quiz1 quiz2 quiz3 quiz4 quiz5 final total percent grade
2      2      1    10    10     7     6     9    54    96      77     C
3      2      2    10     7     8     9     7    57
6      1      2    10    10    10     9     9    74
9      2      2    10    10    10    10     9    74
12     1      2     8    10    10    10     9    71
13     1      1     8     8    10    10     6    69
   passfail
2         P
3         P
6         P
9         P
12        P
13        P
> boxplot(ethnicity_white$final, main =" Box Plot of final for WHITES", col =
"green")
>
```

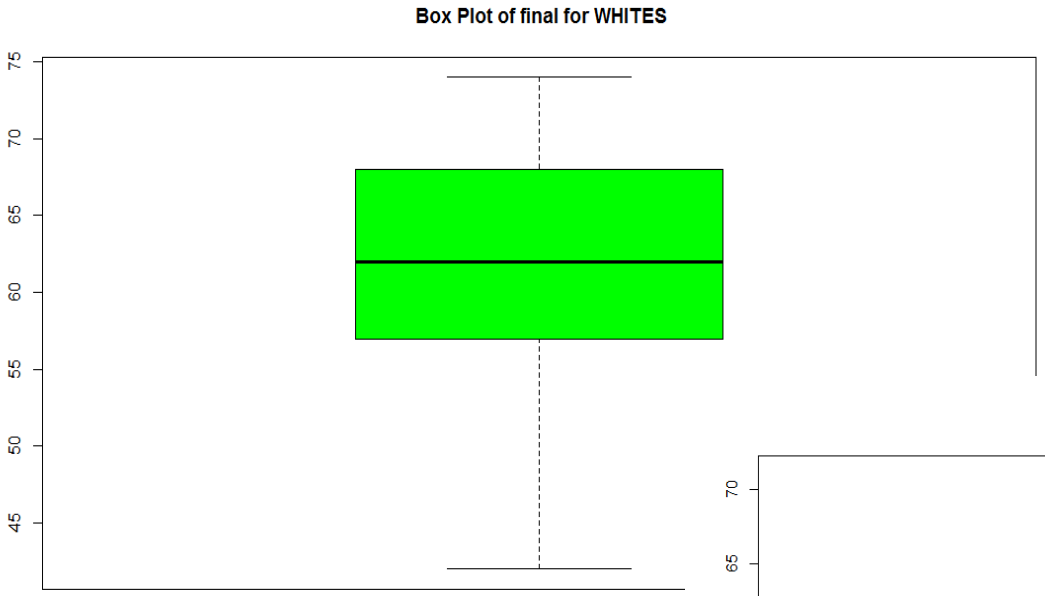


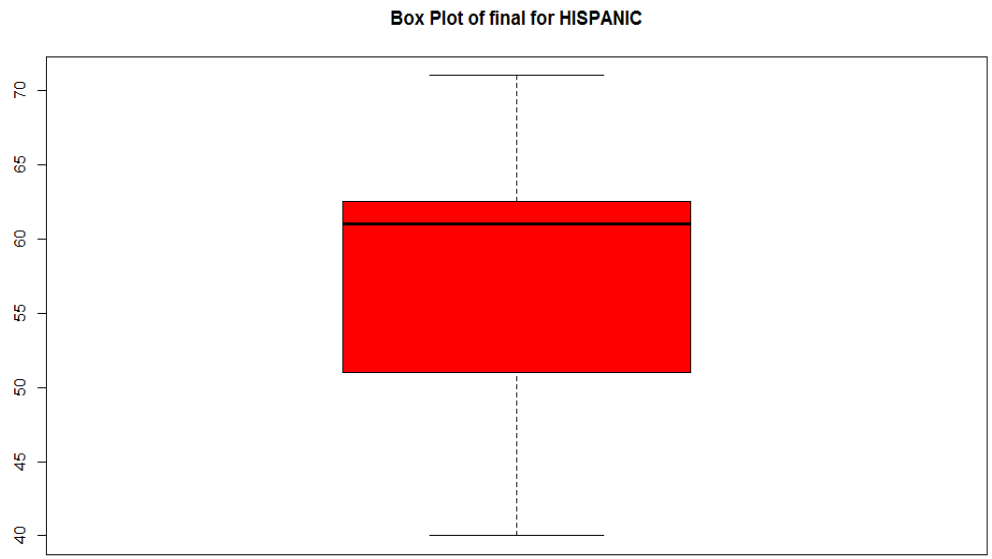
Create subset of only HISPANICS (ethnicity = 5) and view few top rows and make box plot of final

```
> ethnicity_hispanic<-subset(grades, ethnicity == 5)
> head(ethnicity_hispanic)
   Sr_No     id   lastname firstname gender ethnicity year lowup section
8      8 154441       LIAN     JENNY      1         5    2     1       1
16    16 219593     POTTER    MICKEY      1         5    3     2       3
23    23 287617   CUMMINGS    DAVENA      1         5    3     2       3
39    39 447659 GALANVILLE      DANA      1         5    4     2       3
45    45 490016    STEPHEN      LIZA      1         5    3     2       2
46    46 498900      HUANG       JOE      2         5    3     2       3
    gpa extrc review quiz1 quiz2 quiz3 quiz4 quiz5 final total percent grade
8  3.57     1      2    10     9    10    10    10    71   120      96     A
16 2.54     1      2     5     8     6     4    10    61    94      75     C
23 2.21     1      2     9    10     9     9     9    52    98      78     C
39 2.77     1      1     6     8     9     5     8    63    99      79     C
45 2.72     1      2     8     9     9     8    10    60   104      83     B
46 2.47     1      1     0     5     0     2     5    40    52      42     F
   passfail
8         P
16        P
23        P
39        P
45        P
46        F
> boxplot(ethnicity_hispanic$final, main = "Box Plot of final for HISPANIC", c
ol = "red")
>
```



Compare box plots of final for WHITES and HISPANIC





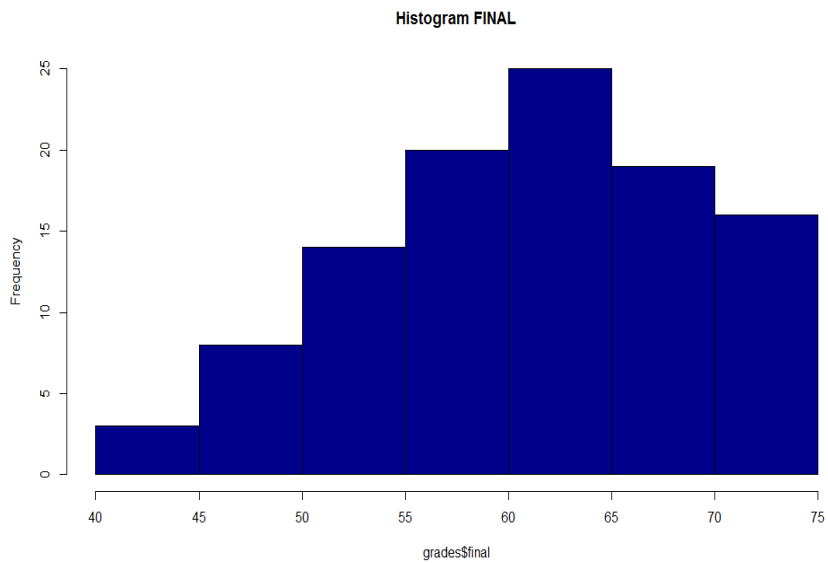
Transform ***final*** with ***square root*** and recode as new variable

```
> grades$sqrtfinal<-sqrt(grades$final)
> head(grades)
  Sr_No     id  lastname firstname gender ethnicity year lowup section  gpa
1     1 106484 VILLARRUZ    ALFRED      2         2    2     1       2 1.18
2     2 108642 VALAZQUEZ     SCOTT      2         4    3     2       2 2.19
3     3 127285    GALVEZ    JACKIE      1         4    4     2       2 2.46
4     4 132931   OSBORNE       ANN      1         3    2     1       2 3.98
5     5 140219    GUADIZ   VALERIE      1         2    4     2       1 1.84
6     6 142630   RANGIFO   TANIECE      1         4    3     2       3 3.90
  extrc review quiz1 quiz2 quiz3 quiz4 quiz5 final total percent grade
1     1      2     6     5     7     6     3    53    80      64     D
2     2      1    10    10     7     6     9    54    96      77     C
3     2      2    10     7     8     9     7    57    98      78     C
4     1      1     7     8     7     7     6    68   103      82     B
5     1      1     7     8     9     8    10    66   108      86     B
6     1      2    10    10    10     9     9    74   122      98     A
  passfail sqrtfinal
1        P  7.280110
2        P  7.348469
3        P  7.549834
4        P  8.246211
5        P  8.124038
6        P  8.602325
>
```

ethnicity_white

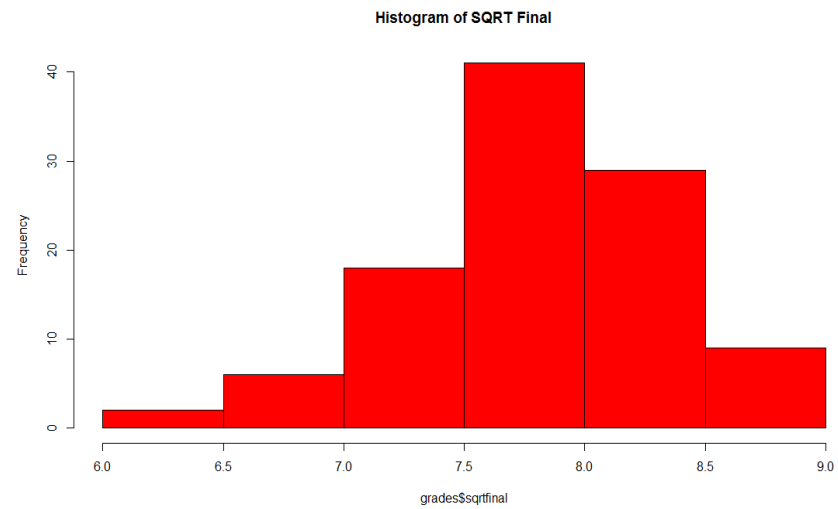
percent	grade	passfail	sqrtfinal
64	D	P	7.280110
77	C	P	7.348469
78	C	P	7.549834
82	B	P	8.246211

Compare Histograms of *final* and *sqrtfinal*



Final: skewness = **-0.33** ; kurtosis = **-0.42**

Sqrtfinal: skewness = **-0.48** ; kurtosis = **-0.17**



Convert *final* into two categories of final
[one, <60 and second >60, 60 will fall in > 60]

```
> grades$catgryfinal<- ifelse(grades$final<60, yes = "final<60", no ="final>60
")
> head(grades)
  Sr_No     id  lastname firstname gender ethnicity year lowup section  gpa
1     1 106484 VILLARRUZ    ALFRED      2         2    2     1       2 1.18
2     2 108642 VALAZQUEZ     SCOTT      2         4    3     2       2 2.19
3     3 127285    GALVEZ    JACKIE      1         4    4     2       2 2.46
4     4 132931   OSBORNE       ANN      1         3    2     1       2 3.98
5     5 140219    GUADIZ   VALERIE      1         2    4     2       1 1.84
6     6 142630   RANGIFO   TANIECE      1         4    3     2       3 3.90
  extrc review quiz1 quiz2 quiz3 quiz4 quiz5 final total percent grade
1     1      2     6     5     7     6     3    53    80      64     D
2     2      1    10    10     7     6     9    54    96      77     C
3     2      2    10     7     8     9     7    57    98      78     C
4     1      1     7     8     7     7     6    68   103      82     B
5     1      1     7     8     9     8    10    66   108      86     B
6     1      2    10    10    10     9     9    74   122      98     A
  passfail sqrtfinal catgryfinal
1        P  7.280110    final<60
2        P  7.348469    final<60
3        P  7.549834    final<60
4        P  8.246211    final>60
5        P  8.124038    final>60
6        P  8.602325    final>60
> table(grades$catgryfinal)

final<60 final>60
      38       67
```

Convert final into two categories of final [one, <60 and second >60]

File Edit Code View Plots Session Build Debug Tools Help

Go to file/function

grades | cs2m | final_60 | ethnicity_native | ethnicity_white

Filter

z2	quiz3	quiz4	quiz5	final	total	percent	grade	passfail	sqrtfinal	catgryfinal
5	7	6	3	53	80	64	D	P	7.280110	final<60
10	7	6	9	54	96	77	C	P	7.348469	final<60
7	8	9	7	57	98	78	C	P	7.549834	final<60
8	7	7	6	68	103	82	B	P	8.246211	final>60

Convert *final* into categories with increment of 5

[40-45=1, 46-50=2, 51-55=3, 56-60=4, 61-65=5, 66-70=6, 71-75=7]

```
> grades$final_cat<-cut(grades$final, breaks = seq(40, 75, 5), labels =c("fina
l1", "final2", "final3", "final4", "final5", "final6", "final7"))
> head(grades)
  Sr_No     id  lastname firstname gender ethnicity year lowup section  gpa
1     1 106484 VILLARRUZ    ALFRED      2         2    2     1       2 1.18
2     2 108642 VALAZQUEZ     SCOTT      2         4    3     2       2 2.19
3     3 127285    GALVEZ    JACKIE      1         4    4     2       2 2.46
4     4 132931   OSBORNE       ANN      1         3    2     1       2 3.98
5     5 140219    GUADIZ   VALERIE      1         2    4     2       1 1.84
6     6 142630   RANGIFO   TANIECE      1         4    3     2       3 3.90
  extrc review quiz1 quiz2 quiz3 quiz4 quiz5 final total percent grade
1     1      2     6     5     7     6     3    53    80      64     D
2     2      1    10    10     7     6     9    54    96      77     C
3     2      2    10     7     8     9     7    57    98      78     C
4     1      1     7     8     7     7     6    68   103      82     B
5     1      1     7     8     9     8    10    66   108      86     B
6     1      2    10    10    10     9     9    74   122      98     A
  passfail sqrtfinal catgryfinal final_cat
1        P  7.280110    final<60    final3
2        P  7.348469    final<60    final3
3        P  7.549834    final<60    final4
4        P  8.246211    final>60    final6
5        P  8.124038    final>60    final6
6        P  8.602325    final>60    final7
> table(grades$catgryfinal)

final<60 final>60
      38       67
> table(grades$final_cat)

final1 final2 final3 final4 final5 final6 final7
     2      8     14     20     25     19     16
```

cut command

within()

Create new variable *agecat* as categories of *Age*

within command

```
library(readr)
k<- read_csv("C:/Users/iNurture/Desktop/Data Sets/cs2m.csv")
str(k)
summary(k$Age)
# using within()
m=k
summary(m)
View(m)
```

```
m <- within(m,{
            agecat<- NA
            agecat[Age>=15 & Age <= 25]  <- 'Low'
            agecat[Age>=26 & Age <= 40]  <- 'Middle'
            agecat[Age>41]               <- 'High'
            })
head(m, 3)
```

```r
m <- within(m,{
            agecat<- NA
            agecat[Age>=15 & Age <= 25]  <- 'Low'
            agecat[Age>=26 & Age <= 40]  <- 'Middle'
            agecat[Age>41]               <- 'High'
            })
head(m, 3)
```

	BP	Chlstrl	Age	Prgnt	AnxtyLH	DrugR	agecat
1	100	150	20	0	0	0	Low
2	120	160	16	0	0	0	Low
3	110	150	18	0	0	0	Low
4	100	175	25	0	0	0	Low
5	95	250	36	0	0	0	Middle
6	110	200	56	0	1	0	High
7	120	180	59	0	1	0	High
8	150	175	45	0	1	0	High
9	160	185	40	0	1	0	Middle
10	125	195	20	1	0	0	Low

Converting *ethnicity* into two categories [category 1= 1, 3& 5; category 2 = 2 & 4

```
> grades$cateth<-grades$ethnicity
> grades$cateth[grades$cateth == 1|grades$cateth == 3|grades$cateth  ==
5]=1
> grades$cateth[grades$cateth == 2|grades$cateth == 4] = 2
```

```
  Sr_No     id  lastname firstname gender ethnicity year lowup section  gpa
1     1 106484 VILLARRUZ    ALFRED      2         2    2     1       2
1.18
2     2 108642 VALAZQUEZ     SCOTT      2         4    3     2       2
2.19
3     3 127285    GALVEZ    JACKIE      1         4    4     2       2
2.46
4     4 132931   OSBORNE       ANN      1         3    2     1       2
3.98
5     5 140219    GUADIZ   VALERIE      1         2    4     2       1
1.84
6     6 142630   RANGIFO   TANIECE      1         4    3     2       3
3.90
  extrc review quiz1 quiz2 quiz3 quiz4 quiz5 final total percent grade
1     1      2     6     5     7     6     3    53    80      64     D
2     2      1    10    10     7     6     9    54    96      77     C
3     2      2    10     7     8     9     7    57    98      78     C
4     1      1     7     8     7     7     6    68   103      82     B
5     1      1     7     8     9     8    10    66   108      86     B
6     1      2    10    10    10     9     9    74   122      98     A
  passfail cateth
1        P      2
2        P      2
3        P      2
4        P      1
5        P      2
6        P      2
```

Take out 20% observations randomly from the file *grades*

```
> sam<-sample(x=1:nrow(grades), size = 0.2*nrow(grades))
> grade20<-grades[sam,]
> head(grade20)
   sr_No     id lastname firstname gender ethnicity year lowup section  gpa
9      9 157147   BAKKEN      KREG      2         4    3     2       1 3.95
44    44 479547 LANGFORD     BLAIR      2         3    3     2       1 3.42
28    28 354601   CARPIO      MARY      1         2    2     1       1 2.03
59    59 616095 SPRINGER  ANNELIES      1         4    3     2       1 3.64
87    87 899529  HAWKINS CARHERINE      1         3    4     2       2 2.31
46    46 498900    HUANG       JOE      2         5    3     2       3 2.47
   extrc review quiz1 quiz2 quiz3 quiz4 quiz5 final total percent grade
9      2      2    10    10    10    10     9    74   123      98     A
44     2      2    10    10    10     9    10    75   124      99     A
28     1      2    10    10    10    10     9    71   120      96     A
59     1      2    10    10    10    10    10    72   122      98     A
87     1      1    10     8     9    10     7    49    93      74     C
46     1      1     0     5     0     2     5    40    52      42     F
   passfail sqrtfinal catgryfinal final_cat
9         P  8.602325    final>60    final7
44        P  8.660254    final>60    final7
28        P  8.426150    final>60    final7
59        P  8.485281    final>60    final7
```

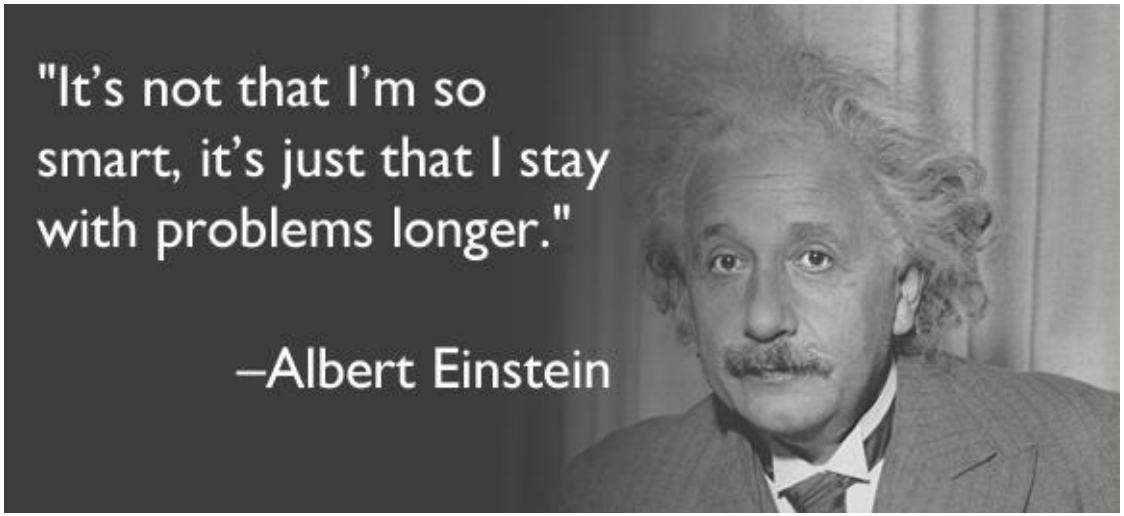
All 20% cases [21#]

```
> grade20
    Sr_No     id   lastname  firstname gender ethnicity year lowup section
9       9 157147     BAKKEN       KREG      2         4    3     2       1
44     44 479547   LANGFORD      BLAIR      2         3    3     2       1
28     28 354601     CARPIO       MARY      1         2    2     1       1
59     59 616095   SPRINGER   ANNELIES      1         4    3     2       1
87     87 899529    HAWKINS  CARHERINE      1         3    4     2       2
46     46 498900      HUANG        JOE      2         5    3     2       3
36     36 420327     BADGER    SUZANNA      1         4    3     2       3
47     47 506467 SCARBROUGH     CYNTHE      1         4    3     2       2
100   100 973427       ROSS      MARIA      1         4    4     2       1
85     85 897606   GENOBAGA JACQUELINE      1         2    3     2       3
84     84 896972      HUANG      MIRNA      1         2    3     2       1
22     22 280440      CHANG       RENE      1         2    3     2       2
11     11 164842 VALENZUELA      NANCY      1         1    4     2       2
37     37 434571       SURI     MATHEW      2         2    3     2       2
101   101 978889    ZIMCHEK    ARMANDO      2         4    4     2       1
10     10 164605   LANGFORD       DAWN      1         3    3     2       2
81     81 822485 VALENZUELA    KATHRYN      1         4    1     1       1
65     65 721311       SONG       LOIS      2         2    3     2       3
71     71 762813       DAEL       IVAN      2         3    2     1       1
103   103 983522      SLOAT      AARON      2         3    3     2       3
61     61 664653       KHAN       JOHN      2         4    3     2       3
```

Case numbers

```
> sam
 [1]   9  44  28  59  87  46  36  47 100  85  84  22  11  37 101  10  81  65
[19]  71 103  61
>
```


You can do it!


"It's not that I'm so smart, it's just that I stay with problems longer."
–Albert Einstein