In [1]:
```
# importing the tidyverse package to manipulate and use the data as I desired. I wil
install.packages("tidyverse")
```

```
also installing the dependencies 'bit', 'fs', 'rappdirs', 'bit64', 'progress', 'proc
essx', 'xfun', 'blob', 'lifecycle', 'vctrs', 'glue', 'tidyselect', 'data.table', 'ga
rgle', 'ids', 'rematch2', 'isoband', 'cpp11', 'ellipsis', 'vroom', 'tzdb', 'callr',
'knitr', 'withr', 'broom', 'cli', 'crayon', 'dbplyr', 'dplyr', 'dtplyr', 'forcats',
'googledrive', 'googlesheets4', 'ggplot2', 'haven', 'hms', 'httr', 'jsonlite', 'lubr
idate', 'magrittr', 'modelr', 'pillar', 'purrr', 'readr', 'reprex', 'rlang', 'rstudi
oapi', 'rvest', 'tibble', 'tidyr', 'xml2'

  There are binary versions available but the source versions are later:
             binary source needs_compilation
fs            1.5.0  1.5.2              TRUE
xfun           0.22   0.28              TRUE
blob          1.2.1  1.2.2             FALSE
lifecycle     1.0.0  1.0.1             FALSE
glue          1.4.2  1.5.1              TRUE
data.table   1.14.0 1.14.2              TRUE
gargle        1.1.0  1.2.0             FALSE
isoband       0.2.4  0.2.5              TRUE
cpp11         0.2.7  0.4.2             FALSE
vroom         1.4.0  1.5.7              TRUE
tzdb          0.1.1  0.2.0              TRUE
knitr          1.33   1.36             FALSE
withr         2.4.2  2.4.3             FALSE
broom         0.7.6 0.7.10             FALSE
cli           2.5.0  3.1.0              TRUE
crayon        1.4.1  1.4.2             FALSE
dplyr         1.0.6  1.0.7              TRUE
dtplyr        1.1.0  1.2.0             FALSE
googledrive   1.0.1  2.0.0             FALSE
googlesheets4 0.3.0  1.0.0             FALSE
ggplot2       3.3.3  3.3.5             FALSE
haven         2.4.1  2.4.3              TRUE
hms           1.0.0  1.1.1             FALSE
lubridate    1.7.10  1.8.0              TRUE
pillar        1.6.0  1.6.4             FALSE
readr         1.4.0  2.1.1              TRUE
reprex        2.0.0  2.0.1             FALSE
rlang        0.4.11 0.4.12              TRUE
rvest         1.0.0  1.0.2             FALSE
tibble        3.1.1  3.1.6              TRUE
tidyr         1.1.3  1.1.4              TRUE
xml2          1.3.2  1.3.3              TRUE

  Binaries will be installed
package 'bit' successfully unpacked and MD5 sums checked
package 'fs' successfully unpacked and MD5 sums checked
package 'rappdirs' successfully unpacked and MD5 sums checked
package 'bit64' successfully unpacked and MD5 sums checked
package 'progress' successfully unpacked and MD5 sums checked
package 'processx' successfully unpacked and MD5 sums checked
package 'xfun' successfully unpacked and MD5 sums checked
package 'vctrs' successfully unpacked and MD5 sums checked
package 'glue' successfully unpacked and MD5 sums checked
package 'tidyselect' successfully unpacked and MD5 sums checked
package 'data.table' successfully unpacked and MD5 sums checked
package 'ids' successfully unpacked and MD5 sums checked
package 'rematch2' successfully unpacked and MD5 sums checked
package 'isoband' successfully unpacked and MD5 sums checked
package 'ellipsis' successfully unpacked and MD5 sums checked
```

```
package 'vroom' successfully unpacked and MD5 sums checked
package 'tzdb' successfully unpacked and MD5 sums checked
package 'callr' successfully unpacked and MD5 sums checked
package 'cli' successfully unpacked and MD5 sums checked
package 'dbplyr' successfully unpacked and MD5 sums checked
package 'dplyr' successfully unpacked and MD5 sums checked
package 'forcats' successfully unpacked and MD5 sums checked
package 'haven' successfully unpacked and MD5 sums checked
package 'httr' successfully unpacked and MD5 sums checked
package 'jsonlite' successfully unpacked and MD5 sums checked
```

```
Warning message:
"cannot remove prior installation of package 'jsonlite'"Warning message in file.copy
(savedcopy, lib, recursive = TRUE):
"problem copying C:\Institutions\Gre\Apps\Anaconda3-2021.05\envs\r\Lib\R\library\00L
OCK\jsonlite\libs\x64\jsonlite.dll to C:\Institutions\Gre\Apps\Anaconda3-2021.05\env
s\r\Lib\R\library\jsonlite\libs\x64\jsonlite.dll: Permission denied"Warning message:
"restored 'jsonlite'"
package 'lubridate' successfully unpacked and MD5 sums checked
package 'magrittr' successfully unpacked and MD5 sums checked
package 'modelr' successfully unpacked and MD5 sums checked
package 'purrr' successfully unpacked and MD5 sums checked
package 'readr' successfully unpacked and MD5 sums checked
package 'rlang' successfully unpacked and MD5 sums checked
package 'rstudioapi' successfully unpacked and MD5 sums checked
package 'tibble' successfully unpacked and MD5 sums checked
package 'tidyr' successfully unpacked and MD5 sums checked
package 'xml2' successfully unpacked and MD5 sums checked
package 'tidyverse' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\io6627a\AppData\Local\Temp\Rtmp6B9fsT\downloaded_packages
```

```
installing the source packages 'blob', 'lifecycle', 'gargle', 'cpp11', 'knitr', 'wit
hr', 'broom', 'crayon', 'dtplyr', 'googledrive', 'googlesheets4', 'ggplot2', 'hms',
'pillar', 'reprex', 'rvest'

Warning message in install.packages("tidyverse"):
"installation of package 'dtplyr' had non-zero exit status"Warning message in instal
l.packages("tidyverse"):
"installation of package 'ggplot2' had non-zero exit status"
```

In [2]:
```
library(tidyr)
```

```
Warning message:
"package 'tidyr' was built under R version 3.6.3"
```

In [3]:
```
# initiliazing dataset to df and i will use it
df<-tidyr::who
```

In [4]:
```
# displaying the dataset to check
tidyr::who
```

country	iso2	iso3	year	new_sp_m014	new_sp_m1524	new_sp_m2534	new_sp_m3544	new_sp
Afghanistan	AF	AFG	1980	NA	NA	NA	NA	
Afghanistan	AF	AFG	1981	NA	NA	NA	NA	
Afghanistan	AF	AFG	1982	NA	NA	NA	NA	
Afghanistan	AF	AFG	1983	NA	NA	NA	NA	

country	iso2	iso3	year	new_sp_m014	new_sp_m1524	new_sp_m2534	new_sp_m3544	new_sp
Afghanistan	AF	AFG	1984	NA	NA	NA	NA	
Afghanistan	AF	AFG	1985	NA	NA	NA	NA	
Afghanistan	AF	AFG	1986	NA	NA	NA	NA	
Afghanistan	AF	AFG	1987	NA	NA	NA	NA	
Afghanistan	AF	AFG	1988	NA	NA	NA	NA	
Afghanistan	AF	AFG	1989	NA	NA	NA	NA	
Afghanistan	AF	AFG	1990	NA	NA	NA	NA	
Afghanistan	AF	AFG	1991	NA	NA	NA	NA	
Afghanistan	AF	AFG	1992	NA	NA	NA	NA	
Afghanistan	AF	AFG	1993	NA	NA	NA	NA	
Afghanistan	AF	AFG	1994	NA	NA	NA	NA	
Afghanistan	AF	AFG	1995	NA	NA	NA	NA	
Afghanistan	AF	AFG	1996	NA	NA	NA	NA	
Afghanistan	AF	AFG	1997	0	10	6	3	
Afghanistan	AF	AFG	1998	30	129	128	90	
Afghanistan	AF	AFG	1999	8	55	55	47	
Afghanistan	AF	AFG	2000	52	228	183	149	
Afghanistan	AF	AFG	2001	129	379	349	274	
Afghanistan	AF	AFG	2002	90	476	481	368	
Afghanistan	AF	AFG	2003	127	511	436	284	
Afghanistan	AF	AFG	2004	139	537	568	360	
Afghanistan	AF	AFG	2005	151	606	560	472	
Afghanistan	AF	AFG	2006	193	837	791	574	
Afghanistan	AF	AFG	2007	186	856	840	597	
Afghanistan	AF	AFG	2008	187	941	773	545	
Afghanistan	AF	AFG	2009	200	906	705	499	
...	...	...	...	...	...	...	...	
Zimbabwe	ZW	ZWE	1984	NA	NA	NA	NA	
Zimbabwe	ZW	ZWE	1985	NA	NA	NA	NA	
Zimbabwe	ZW	ZWE	1986	NA	NA	NA	NA	
Zimbabwe	ZW	ZWE	1987	NA	NA	NA	NA	
Zimbabwe	ZW	ZWE	1988	NA	NA	NA	NA	
Zimbabwe	ZW	ZWE	1989	NA	NA	NA	NA	
Zimbabwe	ZW	ZWE	1990	NA	NA	NA	NA	
Zimbabwe	ZW	ZWE	1991	NA	NA	NA	NA	
Zimbabwe	ZW	ZWE	1992	NA	NA	NA	NA	

country	iso2	iso3	year	new_sp_m014	new_sp_m1524	new_sp_m2534	new_sp_m3544	new_sp
Zimbabwe	ZW	ZWE	1993	NA	NA	NA	NA	
Zimbabwe	ZW	ZWE	1994	NA	NA	NA	NA	
Zimbabwe	ZW	ZWE	1995	NA	NA	NA	NA	
Zimbabwe	ZW	ZWE	1996	NA	NA	NA	NA	
Zimbabwe	ZW	ZWE	1997	NA	NA	NA	NA	
Zimbabwe	ZW	ZWE	1998	NA	NA	NA	NA	
Zimbabwe	ZW	ZWE	1999	NA	NA	NA	NA	
Zimbabwe	ZW	ZWE	2000	NA	NA	NA	NA	
Zimbabwe	ZW	ZWE	2001	NA	NA	NA	NA	
Zimbabwe	ZW	ZWE	2002	191	600	2548	1662	
Zimbabwe	ZW	ZWE	2003	133	874	3048	2228	
Zimbabwe	ZW	ZWE	2004	187	833	2908	2298	
Zimbabwe	ZW	ZWE	2005	210	837	2264	1855	
Zimbabwe	ZW	ZWE	2006	215	736	2391	1939	
Zimbabwe	ZW	ZWE	2007	138	500	3693	0	
Zimbabwe	ZW	ZWE	2008	127	614	0	3316	
Zimbabwe	ZW	ZWE	2009	125	578	NA	3471	
Zimbabwe	ZW	ZWE	2010	150	710	2208	1682	
Zimbabwe	ZW	ZWE	2011	152	784	2467	2071	
Zimbabwe	ZW	ZWE	2012	120	783	2421	2086	
Zimbabwe	ZW	ZWE	2013	NA	NA	NA	NA	

In [5]:

```
## Q.1
##Gather together all the columns from _new_spm014 to _newrelf65
##Use pivot_longer()
## We do not know what those values represent yet, so we give them the
## generic name as "key"
## We know the cells represent the count of cases, so we use the variable cases.
## There are a lot of missing values in the current representation, so for now we
## use values_drop_na just so we can focus on the values that are present.

# between 'new_sp_m014' and 'newrel_f65' all the values are Nan and we connected tho
# we created key and cases coluns and droppend the NaN values using values_drop_na =


who1<-who %>% pivot_longer(c('new_sp_m014':'newrel_f65'),
  names_to = "key",
  values_to = "cases",
  values_drop_na = TRUE
)
```

In [6]:

```
# displaying the dataset to check
who1
```

country	iso2	iso3	year	key	cases
Afghanistan	AF	AFG	1997	new_sp_m014	0
Afghanistan	AF	AFG	1997	new_sp_m1524	10
Afghanistan	AF	AFG	1997	new_sp_m2534	6
Afghanistan	AF	AFG	1997	new_sp_m3544	3
Afghanistan	AF	AFG	1997	new_sp_m4554	5
Afghanistan	AF	AFG	1997	new_sp_m5564	2
Afghanistan	AF	AFG	1997	new_sp_m65	0
Afghanistan	AF	AFG	1997	new_sp_f014	5
Afghanistan	AF	AFG	1997	new_sp_f1524	38
Afghanistan	AF	AFG	1997	new_sp_f2534	36
Afghanistan	AF	AFG	1997	new_sp_f3544	14
Afghanistan	AF	AFG	1997	new_sp_f4554	8
Afghanistan	AF	AFG	1997	new_sp_f5564	0
Afghanistan	AF	AFG	1997	new_sp_f65	1
Afghanistan	AF	AFG	1998	new_sp_m014	30
Afghanistan	AF	AFG	1998	new_sp_m1524	129
Afghanistan	AF	AFG	1998	new_sp_m2534	128
Afghanistan	AF	AFG	1998	new_sp_m3544	90
Afghanistan	AF	AFG	1998	new_sp_m4554	89
Afghanistan	AF	AFG	1998	new_sp_m5564	64
Afghanistan	AF	AFG	1998	new_sp_m65	41
Afghanistan	AF	AFG	1998	new_sp_f014	45
Afghanistan	AF	AFG	1998	new_sp_f1524	350
Afghanistan	AF	AFG	1998	new_sp_f2534	419
Afghanistan	AF	AFG	1998	new_sp_f3544	194
Afghanistan	AF	AFG	1998	new_sp_f4554	118
Afghanistan	AF	AFG	1998	new_sp_f5564	61
Afghanistan	AF	AFG	1998	new_sp_f65	20
Afghanistan	AF	AFG	1999	new_sp_m014	8
Afghanistan	AF	AFG	1999	new_sp_m1524	55
...	...	...	...	...	...
Zimbabwe	ZW	ZWE	2012	new_sn_f5564	516
Zimbabwe	ZW	ZWE	2012	new_sn_f65	432
Zimbabwe	ZW	ZWE	2012	new_ep_m014	233
Zimbabwe	ZW	ZWE	2012	new_ep_m1524	214
Zimbabwe	ZW	ZWE	2012	new_ep_m2534	658

country	iso2	iso3	year	key	cases
Zimbabwe	ZW	ZWE	2012	new_ep_m3544	789
Zimbabwe	ZW	ZWE	2012	new_ep_m4554	331
Zimbabwe	ZW	ZWE	2012	new_ep_m5564	178
Zimbabwe	ZW	ZWE	2012	new_ep_m65	182
Zimbabwe	ZW	ZWE	2012	new_ep_f014	208
Zimbabwe	ZW	ZWE	2012	new_ep_f1524	319
Zimbabwe	ZW	ZWE	2012	new_ep_f2534	710
Zimbabwe	ZW	ZWE	2012	new_ep_f3544	579
Zimbabwe	ZW	ZWE	2012	new_ep_f4554	228
Zimbabwe	ZW	ZWE	2012	new_ep_f5564	140
Zimbabwe	ZW	ZWE	2012	new_ep_f65	143
Zimbabwe	ZW	ZWE	2013	newrel_m014	1315
Zimbabwe	ZW	ZWE	2013	newrel_m1524	1642
Zimbabwe	ZW	ZWE	2013	newrel_m2534	5331
Zimbabwe	ZW	ZWE	2013	newrel_m3544	5363
Zimbabwe	ZW	ZWE	2013	newrel_m4554	2349
Zimbabwe	ZW	ZWE	2013	newrel_m5564	1206
Zimbabwe	ZW	ZWE	2013	newrel_m65	1208
Zimbabwe	ZW	ZWE	2013	newrel_f014	1252
Zimbabwe	ZW	ZWE	2013	newrel_f1524	2069
Zimbabwe	ZW	ZWE	2013	newrel_f2534	4649
Zimbabwe	ZW	ZWE	2013	newrel_f3544	3526
Zimbabwe	ZW	ZWE	2013	newrel_f4554	1453
Zimbabwe	ZW	ZWE	2013	newrel_f5564	811
Zimbabwe	ZW	ZWE	2013	newrel_f65	725

In [7]:
```
## Q2. Make variable names consistent
## Instead of _newrel we have newrel. It is hard to spot this here but if you do not
## we will get errors in subsequent steps.
## Use stringr::str_replace() in strings: replace the characters “newrel” with
## “new_rel”.
## Name the dataset who2
```

In [8]:
```
## I used stringr library to use str_replace and dplyr for mutate and every new_rel
## and assigned it to who2 dataset.
library(stringr)
library(dplyr)
who2 <- who1 %>%
  mutate(key = stringr::str_replace(who1$key, "newrel", "new_rel"))
```

```
Warning message:
```

```
"package 'dplyr' was built under R version 3.6.3"
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

In [9]:

```
# displaying the dataset to check
who2
```

country	iso2	iso3	year	key	cases
Afghanistan	AF	AFG	1997	new_sp_m014	0
Afghanistan	AF	AFG	1997	new_sp_m1524	10
Afghanistan	AF	AFG	1997	new_sp_m2534	6
Afghanistan	AF	AFG	1997	new_sp_m3544	3
Afghanistan	AF	AFG	1997	new_sp_m4554	5
Afghanistan	AF	AFG	1997	new_sp_m5564	2
Afghanistan	AF	AFG	1997	new_sp_m65	0
Afghanistan	AF	AFG	1997	new_sp_f014	5
Afghanistan	AF	AFG	1997	new_sp_f1524	38
Afghanistan	AF	AFG	1997	new_sp_f2534	36
Afghanistan	AF	AFG	1997	new_sp_f3544	14
Afghanistan	AF	AFG	1997	new_sp_f4554	8
Afghanistan	AF	AFG	1997	new_sp_f5564	0
Afghanistan	AF	AFG	1997	new_sp_f65	1
Afghanistan	AF	AFG	1998	new_sp_m014	30
Afghanistan	AF	AFG	1998	new_sp_m1524	129
Afghanistan	AF	AFG	1998	new_sp_m2534	128
Afghanistan	AF	AFG	1998	new_sp_m3544	90
Afghanistan	AF	AFG	1998	new_sp_m4554	89
Afghanistan	AF	AFG	1998	new_sp_m5564	64
Afghanistan	AF	AFG	1998	new_sp_m65	41
Afghanistan	AF	AFG	1998	new_sp_f014	45
Afghanistan	AF	AFG	1998	new_sp_f1524	350
Afghanistan	AF	AFG	1998	new_sp_f2534	419
Afghanistan	AF	AFG	1998	new_sp_f3544	194
Afghanistan	AF	AFG	1998	new_sp_f4554	118
Afghanistan	AF	AFG	1998	new_sp_f5564	61

country	iso2	iso3	year	key	cases
Afghanistan	AF	AFG	1998	new_sp_f65	20
Afghanistan	AF	AFG	1999	new_sp_m014	8
Afghanistan	AF	AFG	1999	new_sp_m1524	55
...	...	...	...	...	...
Zimbabwe	ZW	ZWE	2012	new_sn_f5564	516
Zimbabwe	ZW	ZWE	2012	new_sn_f65	432
Zimbabwe	ZW	ZWE	2012	new_ep_m014	233
Zimbabwe	ZW	ZWE	2012	new_ep_m1524	214
Zimbabwe	ZW	ZWE	2012	new_ep_m2534	658
Zimbabwe	ZW	ZWE	2012	new_ep_m3544	789
Zimbabwe	ZW	ZWE	2012	new_ep_m4554	331
Zimbabwe	ZW	ZWE	2012	new_ep_m5564	178
Zimbabwe	ZW	ZWE	2012	new_ep_m65	182
Zimbabwe	ZW	ZWE	2012	new_ep_f014	208
Zimbabwe	ZW	ZWE	2012	new_ep_f1524	319
Zimbabwe	ZW	ZWE	2012	new_ep_f2534	710
Zimbabwe	ZW	ZWE	2012	new_ep_f3544	579
Zimbabwe	ZW	ZWE	2012	new_ep_f4554	228
Zimbabwe	ZW	ZWE	2012	new_ep_f5564	140
Zimbabwe	ZW	ZWE	2012	new_ep_f65	143
Zimbabwe	ZW	ZWE	2013	new_rel_m014	1315
Zimbabwe	ZW	ZWE	2013	new_rel_m1524	1642
Zimbabwe	ZW	ZWE	2013	new_rel_m2534	5331
Zimbabwe	ZW	ZWE	2013	new_rel_m3544	5363
Zimbabwe	ZW	ZWE	2013	new_rel_m4554	2349
Zimbabwe	ZW	ZWE	2013	new_rel_m5564	1206
Zimbabwe	ZW	ZWE	2013	new_rel_m65	1208
Zimbabwe	ZW	ZWE	2013	new_rel_f014	1252
Zimbabwe	ZW	ZWE	2013	new_rel_f1524	2069
Zimbabwe	ZW	ZWE	2013	new_rel_f2534	4649
Zimbabwe	ZW	ZWE	2013	new_rel_f3544	3526
Zimbabwe	ZW	ZWE	2013	new_rel_f4554	1453
Zimbabwe	ZW	ZWE	2013	new_rel_f5564	811
Zimbabwe	ZW	ZWE	2013	new_rel_f65	725

In [10]:
```
## I runned the following provided command and assigned to who3 and sperated key col
## new,type and sexage columns.
```

```
who3<-who2 %>%
separate(key,c("new","type","sexage"),sep="_")
## purpose of %>%
## It is called pipe operator
## It takes the output of one function and passes it into another function as an arg
## This operator will forward a value or a result of an expression, into the next fu
```

In [11]:

```
# displaying the dataset to check
who3
```

country	iso2	iso3	year	new	type	sexage	cases
Afghanistan	AF	AFG	1997	new	sp	m014	0
Afghanistan	AF	AFG	1997	new	sp	m1524	10
Afghanistan	AF	AFG	1997	new	sp	m2534	6
Afghanistan	AF	AFG	1997	new	sp	m3544	3
Afghanistan	AF	AFG	1997	new	sp	m4554	5
Afghanistan	AF	AFG	1997	new	sp	m5564	2
Afghanistan	AF	AFG	1997	new	sp	m65	0
Afghanistan	AF	AFG	1997	new	sp	f014	5
Afghanistan	AF	AFG	1997	new	sp	f1524	38
Afghanistan	AF	AFG	1997	new	sp	f2534	36
Afghanistan	AF	AFG	1997	new	sp	f3544	14
Afghanistan	AF	AFG	1997	new	sp	f4554	8
Afghanistan	AF	AFG	1997	new	sp	f5564	0
Afghanistan	AF	AFG	1997	new	sp	f65	1
Afghanistan	AF	AFG	1998	new	sp	m014	30
Afghanistan	AF	AFG	1998	new	sp	m1524	129
Afghanistan	AF	AFG	1998	new	sp	m2534	128
Afghanistan	AF	AFG	1998	new	sp	m3544	90
Afghanistan	AF	AFG	1998	new	sp	m4554	89
Afghanistan	AF	AFG	1998	new	sp	m5564	64
Afghanistan	AF	AFG	1998	new	sp	m65	41
Afghanistan	AF	AFG	1998	new	sp	f014	45
Afghanistan	AF	AFG	1998	new	sp	f1524	350
Afghanistan	AF	AFG	1998	new	sp	f2534	419
Afghanistan	AF	AFG	1998	new	sp	f3544	194
Afghanistan	AF	AFG	1998	new	sp	f4554	118
Afghanistan	AF	AFG	1998	new	sp	f5564	61
Afghanistan	AF	AFG	1998	new	sp	f65	20
Afghanistan	AF	AFG	1999	new	sp	m014	8

country	iso2	iso3	year	new	type	sexage	cases
Afghanistan	AF	AFG	1999	new	sp	m1524	55
...	...	...	...	...	...	...	...
Zimbabwe	ZW	ZWE	2012	new	sn	f5564	516
Zimbabwe	ZW	ZWE	2012	new	sn	f65	432
Zimbabwe	ZW	ZWE	2012	new	ep	m014	233
Zimbabwe	ZW	ZWE	2012	new	ep	m1524	214
Zimbabwe	ZW	ZWE	2012	new	ep	m2534	658
Zimbabwe	ZW	ZWE	2012	new	ep	m3544	789
Zimbabwe	ZW	ZWE	2012	new	ep	m4554	331
Zimbabwe	ZW	ZWE	2012	new	ep	m5564	178
Zimbabwe	ZW	ZWE	2012	new	ep	m65	182
Zimbabwe	ZW	ZWE	2012	new	ep	f014	208
Zimbabwe	ZW	ZWE	2012	new	ep	f1524	319
Zimbabwe	ZW	ZWE	2012	new	ep	f2534	710
Zimbabwe	ZW	ZWE	2012	new	ep	f3544	579
Zimbabwe	ZW	ZWE	2012	new	ep	f4554	228
Zimbabwe	ZW	ZWE	2012	new	ep	f5564	140
Zimbabwe	ZW	ZWE	2012	new	ep	f65	143
Zimbabwe	ZW	ZWE	2013	new	rel	m014	1315
Zimbabwe	ZW	ZWE	2013	new	rel	m1524	1642
Zimbabwe	ZW	ZWE	2013	new	rel	m2534	5331
Zimbabwe	ZW	ZWE	2013	new	rel	m3544	5363
Zimbabwe	ZW	ZWE	2013	new	rel	m4554	2349
Zimbabwe	ZW	ZWE	2013	new	rel	m5564	1206
Zimbabwe	ZW	ZWE	2013	new	rel	m65	1208
Zimbabwe	ZW	ZWE	2013	new	rel	f014	1252
Zimbabwe	ZW	ZWE	2013	new	rel	f1524	2069
Zimbabwe	ZW	ZWE	2013	new	rel	f2534	4649
Zimbabwe	ZW	ZWE	2013	new	rel	f3544	3526
Zimbabwe	ZW	ZWE	2013	new	rel	f4554	1453
Zimbabwe	ZW	ZWE	2013	new	rel	f5564	811
Zimbabwe	ZW	ZWE	2013	new	rel	f65	725

In [12]:

```
## q4. Separate sexage into sex and age: Use the function separate(). Name the
##dataset who4

who4<-who3 %>%
separate(sexage,c("sex","age"),sep="(?<=[A-Za-z])(?=[0-9])")
```

```
## I used regex for number and letters. In this regex first we take the letter [A-Za
## ?<= it matches look behind the cursor
## ?= it means check the after the letters as numbers
## I did not put any cursor commands because I wanted to split them as number and le
```

In [13]:

```
## Displaying dataset to check
who4
```

country	iso2	iso3	year	new	type	sex	age	cases
Afghanistan	AF	AFG	1997	new	sp	m	014	0
Afghanistan	AF	AFG	1997	new	sp	m	1524	10
Afghanistan	AF	AFG	1997	new	sp	m	2534	6
Afghanistan	AF	AFG	1997	new	sp	m	3544	3
Afghanistan	AF	AFG	1997	new	sp	m	4554	5
Afghanistan	AF	AFG	1997	new	sp	m	5564	2
Afghanistan	AF	AFG	1997	new	sp	m	65	0
Afghanistan	AF	AFG	1997	new	sp	f	014	5
Afghanistan	AF	AFG	1997	new	sp	f	1524	38
Afghanistan	AF	AFG	1997	new	sp	f	2534	36
Afghanistan	AF	AFG	1997	new	sp	f	3544	14
Afghanistan	AF	AFG	1997	new	sp	f	4554	8
Afghanistan	AF	AFG	1997	new	sp	f	5564	0
Afghanistan	AF	AFG	1997	new	sp	f	65	1
Afghanistan	AF	AFG	1998	new	sp	m	014	30
Afghanistan	AF	AFG	1998	new	sp	m	1524	129
Afghanistan	AF	AFG	1998	new	sp	m	2534	128
Afghanistan	AF	AFG	1998	new	sp	m	3544	90
Afghanistan	AF	AFG	1998	new	sp	m	4554	89
Afghanistan	AF	AFG	1998	new	sp	m	5564	64
Afghanistan	AF	AFG	1998	new	sp	m	65	41
Afghanistan	AF	AFG	1998	new	sp	f	014	45
Afghanistan	AF	AFG	1998	new	sp	f	1524	350
Afghanistan	AF	AFG	1998	new	sp	f	2534	419
Afghanistan	AF	AFG	1998	new	sp	f	3544	194
Afghanistan	AF	AFG	1998	new	sp	f	4554	118
Afghanistan	AF	AFG	1998	new	sp	f	5564	61
Afghanistan	AF	AFG	1998	new	sp	f	65	20
Afghanistan	AF	AFG	1999	new	sp	m	014	8
Afghanistan	AF	AFG	1999	new	sp	m	1524	55

country	iso2	iso3	year	new	type	sex	age	cases
...	...	...	...	...	...	...	...	...
Zimbabwe	ZW	ZWE	2012	new	sn	f	5564	516
Zimbabwe	ZW	ZWE	2012	new	sn	f	65	432
Zimbabwe	ZW	ZWE	2012	new	ep	m	014	233
Zimbabwe	ZW	ZWE	2012	new	ep	m	1524	214
Zimbabwe	ZW	ZWE	2012	new	ep	m	2534	658
Zimbabwe	ZW	ZWE	2012	new	ep	m	3544	789
Zimbabwe	ZW	ZWE	2012	new	ep	m	4554	331
Zimbabwe	ZW	ZWE	2012	new	ep	m	5564	178
Zimbabwe	ZW	ZWE	2012	new	ep	m	65	182
Zimbabwe	ZW	ZWE	2012	new	ep	f	014	208
Zimbabwe	ZW	ZWE	2012	new	ep	f	1524	319
Zimbabwe	ZW	ZWE	2012	new	ep	f	2534	710
Zimbabwe	ZW	ZWE	2012	new	ep	f	3544	579
Zimbabwe	ZW	ZWE	2012	new	ep	f	4554	228
Zimbabwe	ZW	ZWE	2012	new	ep	f	5564	140
Zimbabwe	ZW	ZWE	2012	new	ep	f	65	143
Zimbabwe	ZW	ZWE	2013	new	rel	m	014	1315
Zimbabwe	ZW	ZWE	2013	new	rel	m	1524	1642
Zimbabwe	ZW	ZWE	2013	new	rel	m	2534	5331
Zimbabwe	ZW	ZWE	2013	new	rel	m	3544	5363
Zimbabwe	ZW	ZWE	2013	new	rel	m	4554	2349
Zimbabwe	ZW	ZWE	2013	new	rel	m	5564	1206
Zimbabwe	ZW	ZWE	2013	new	rel	m	65	1208
Zimbabwe	ZW	ZWE	2013	new	rel	f	014	1252
Zimbabwe	ZW	ZWE	2013	new	rel	f	1524	2069
Zimbabwe	ZW	ZWE	2013	new	rel	f	2534	4649
Zimbabwe	ZW	ZWE	2013	new	rel	f	3544	3526
Zimbabwe	ZW	ZWE	2013	new	rel	f	4554	1453
Zimbabwe	ZW	ZWE	2013	new	rel	f	5564	811
Zimbabwe	ZW	ZWE	2013	new	rel	f	65	725

In [14]:

```
## Q.4 Print the first 5 rows and the last 5 rows of the dataset who4 to the screen.
## HEAD function used to print first 5 rows . Also for 5 rows built-in tail function
head(who4, 5)
```

country	iso2	iso3	year	new	type	sex	age	cases

country	iso2	iso3	year	new	type	sex	age	cases
Afghanistan	AF	AFG	1997	new	sp	m	014	0
Afghanistan	AF	AFG	1997	new	sp	m	1524	10
Afghanistan	AF	AFG	1997	new	sp	m	2534	6
Afghanistan	AF	AFG	1997	new	sp	m	3544	3
Afghanistan	AF	AFG	1997	new	sp	m	4554	5

In [15]:
```
## to display last 5 rows I used tail() function.
tail(who4,5)
```

country	iso2	iso3	year	new	type	sex	age	cases
Zimbabwe	ZW	ZWE	2013	new	rel	f	2534	4649
Zimbabwe	ZW	ZWE	2013	new	rel	f	3544	3526
Zimbabwe	ZW	ZWE	2013	new	rel	f	4554	1453
Zimbabwe	ZW	ZWE	2013	new	rel	f	5564	811
Zimbabwe	ZW	ZWE	2013	new	rel	f	65	725

In [16]:
```
## Q.6 Export who4 as an csv file and save it in your local directory.
## write.csv method used with the dataset and file name parameters and saved in my l
write.csv(who4, "who4file.csv")
```

In [17]:
```
library(datasets)
##, we'll use the built-in R data set named "iris"
```

In [18]:
```
##, to see variables and values of the dataset.
str(iris)
```

```
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1
...
```

In [19]:
```
## Using summary to see min,1stquarter median,mean,3rd qurter and max values.
summary(iris)
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
Median :5.800   Median :3.000   Median :4.350   Median :1.300
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
setosa    :50
versicolor:50
virginica :50
```

In [20]:
```
## assigning iris to a dataframe
df<-iris
df1 <- iris
```

In [21]:
```
## 1. Compute the mean, median and mode of sepal length
## Mean value using mean() method
mean(df$Sepal.Length)
```

5.84333333333333

In [22]:
```
## Median using median() method
median(df$Sepal.Length)
```

5.8

In [23]:
```
## Mode
## R does not support mode function so I created a function that finds mode using th
getmode <- function(sepal) {
   uniqsepal <- unique(sepal)
   uniqsepal[which.max(tabulate(match(sepal, uniqsepal)))]
}
```

In [25]:
```
## I assigned the return value to a varible and printed it.
result<- getmode(df$Sepal.Length)
print(result)
```

[1] 5

In [26]:
```
## Q2.Compute how “spread out” the data are. Here you need to calculate the minimum,
##and range of sepal length (2 marks).

##Min value using min method
min(df$Sepal.Length)
```

4.3

In [27]:
```
## Max value using max() method
max(df$Sepal.Length)
```

7.9

In [28]:
```
## range using range()method
range(df$Sepal.Length)
```

1. 4.3
2. 7.9

In [29]:
```
## Q3.Calculate the interquartile (IQR) range of sepal length (1 mark). Use the func
##measure quantiles for the same variable, sepal length, and comment what is the dif
##relation of these two functions regarding the results shown on your screen? (2 mar
```

```
##IQR value
IQR(df$Sepal.Length)
```

1.3

In [30]:
```
## quantile value
quantile(df$Sepal.Length)
```

0%	4.3
25%	5.1
50%	5.8
75%	6.4
100%	7.9

In [31]:
```
## Response to 2. question -> IQR is difference between %75 and %25 percentiles of d
## Quantile shows us the percentage of values below a certain value.
```

In [32]:
```
## Q4. Compute the variance (1 mark) and standard deviation of sepal length (1 mark)
# variance value
var(df$Sepal.Length)
```

0.685693512304251

In [33]:
```
## standart deviation value of sepal length
sd(df$Sepal.Length)
```

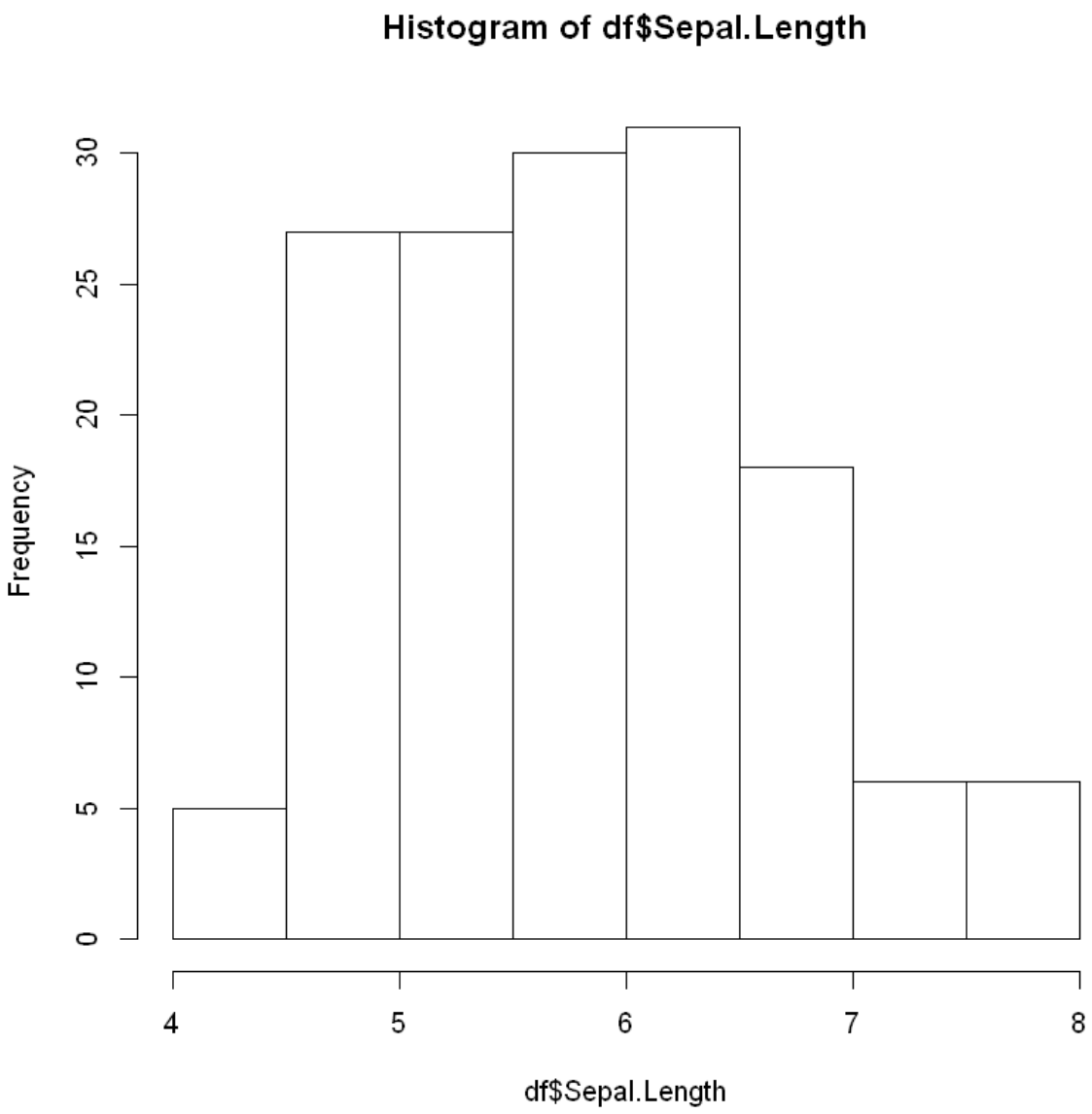
0.828066127977863

In [34]:
```
## Q5.Choose the right function to show min, max, mean, median, 1st and 3rd quantile
##the variable sepal length (1 mark).
fivenum(df$Sepal.Length)
```

1. 4.3
2. 5.1
3. 5.8
4. 6.4
5. 7.9

In [35]:
```
summary(df$Sepal.Length) ## summary now shows the exact desired values.
hist(df$Sepal.Length)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  4.300   5.100   5.800   5.843   6.400   7.900 
```



In [36]:
```
## Q6. Use sapply() to compute the mean (1 marks) and quantiles (1 marks) of each co
## dataset iris.
## quantile values --
sapply(df[, -5], quantile)
```

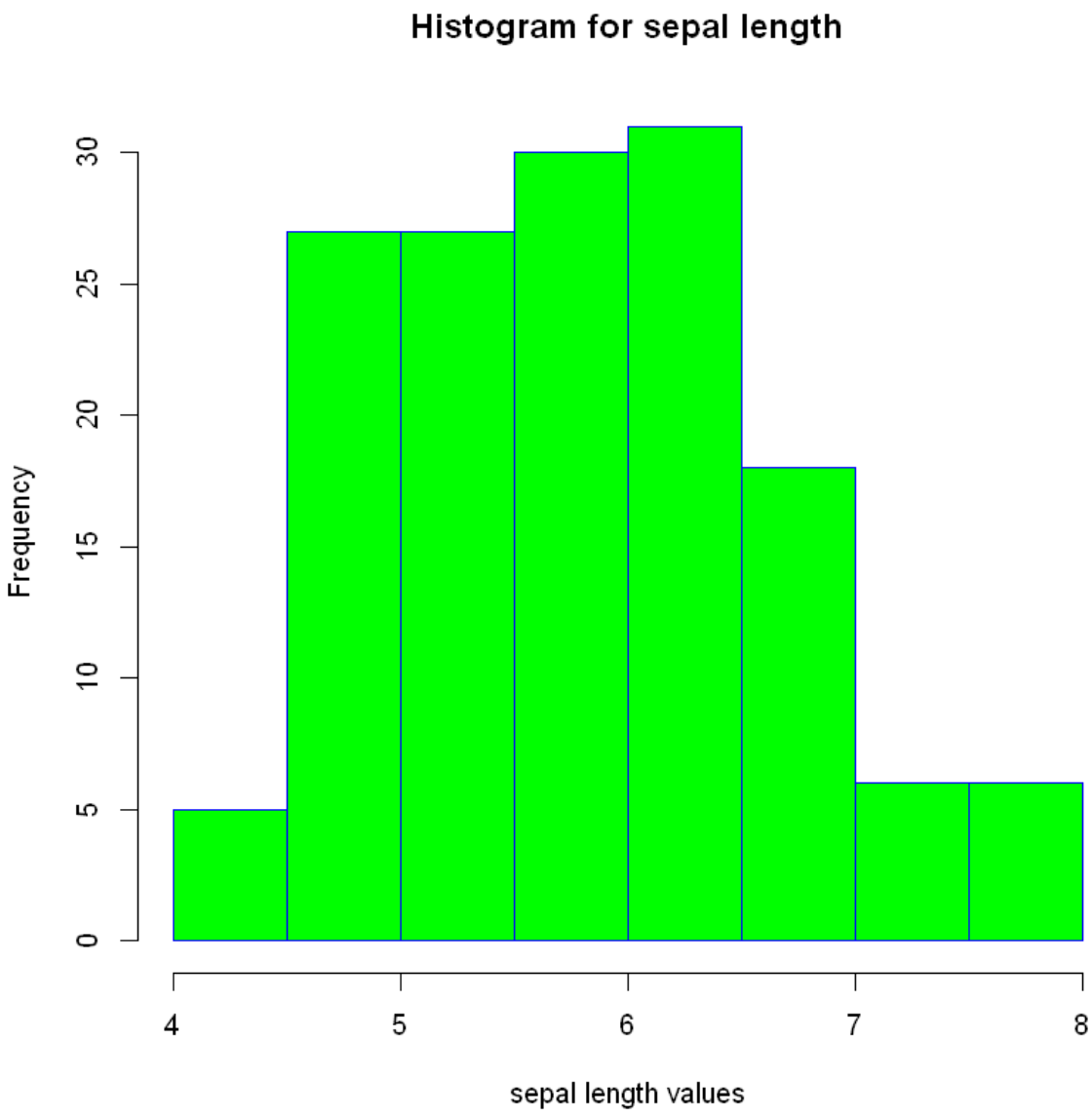
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
0%	4.3	2.0	1.00	0.1
25%	5.1	2.8	1.60	0.3
50%	5.8	3.0	4.35	1.3
75%	6.4	3.3	5.10	1.8
100%	7.9	4.4	6.90	2.5

In [37]:
```
## mean value
sapply(df[, -5], mean)
```

Sepal.Length	5.84333333333333
Sepal.Width	3.05733333333333
Petal.Length	3.758

Petal.Width 1.19933333333333

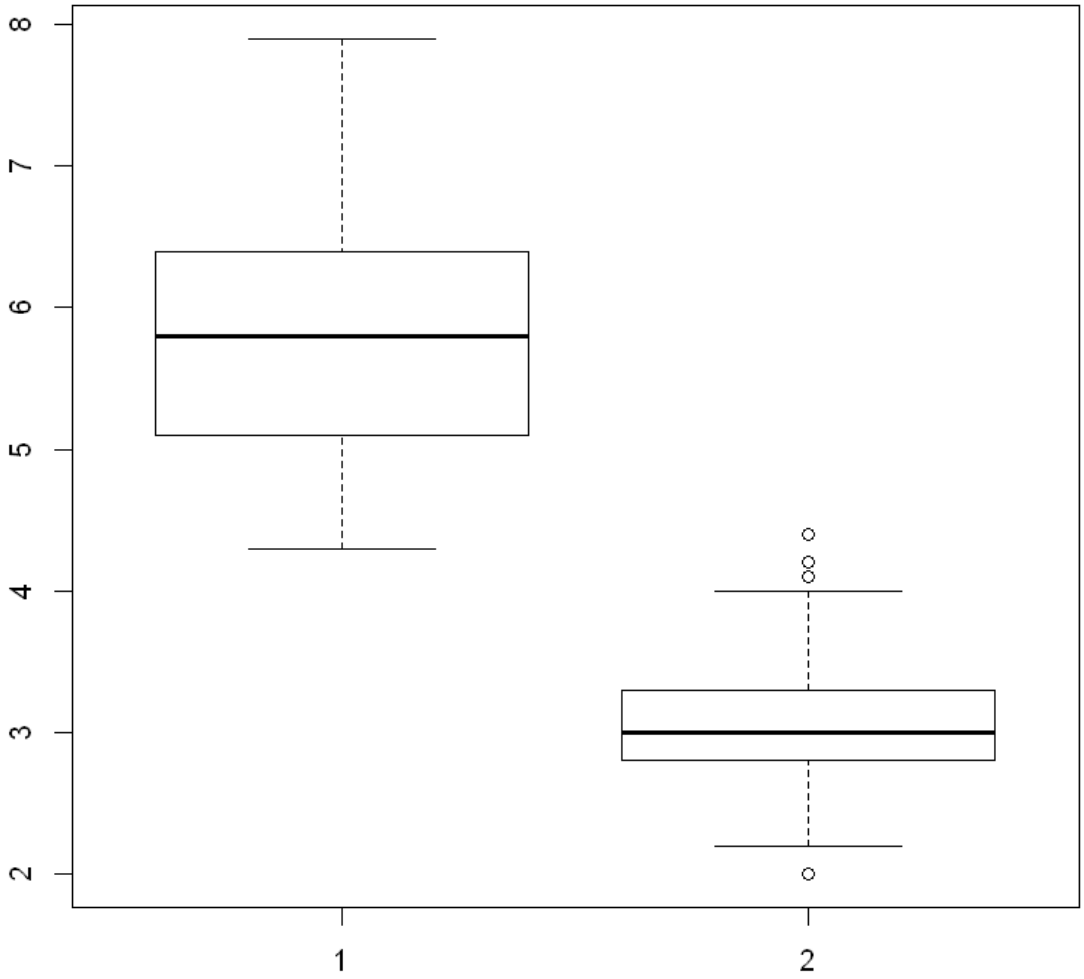
In [38]:
```
## Q7. Use the in-built R basic functions (no need to import any library) to create
##Length. Make sure you add the following arguments: (4 marks)
## main: Add a title for this plot, e.g., “Histogram for sepal length”
## xlab: Add a label for the x axis
## border: Set a colour for the border around the bars, e.g., blue
## col: set a colour of the bars, e.g., green
## used hist method and gave the parameters that mentioned in the question to main,x
## 4 to 8 which is in the sepal length values.
hist(df$Sepal.Length,
     main="Histogram for sepal length",
     xlab="sepal length values",
     border="blue",
     col="green",
     xlim=c(4,8),
     breaks=9)
```



In [39]:
```
##Use the in-built R basic function to create one boxplot for sepal length, sepal wi
## and petal width
```

In [40]:
```
boxplot(df$Sepal.Length,df$Sepal.Width )   #the plotting formula of y ~ x
```

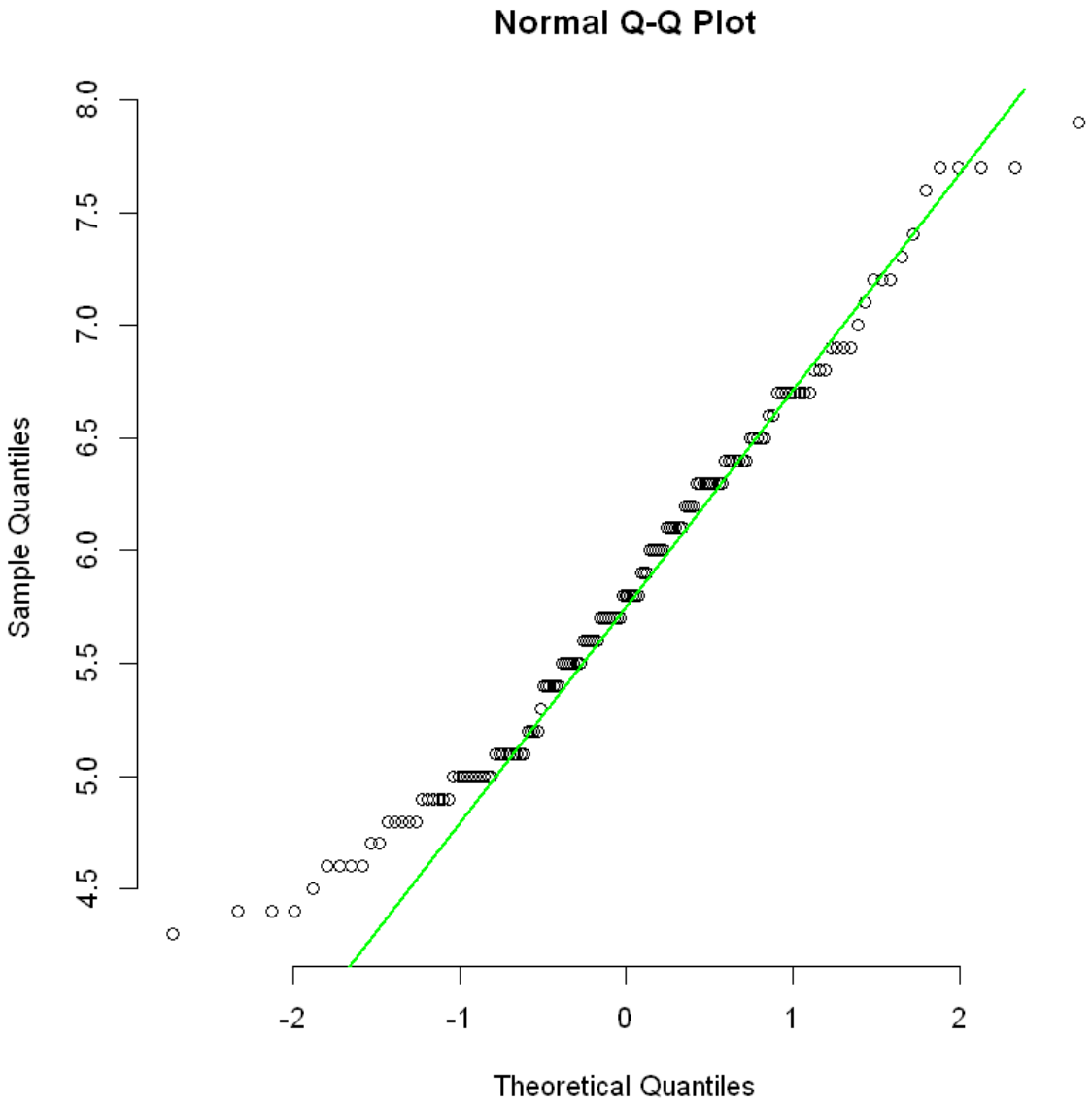
```
#         data = iris,           #the data to plot
#          xlab = "Iris Species", #Label for x axis
#          ylab = "Sepal Length",
#          col = c("red","green","blue"),
#          lwd = 2,
#          cex = 2)
```



In [41]:

```
#  Q.9 The R base functions qqnorm() and qqplot() are used to produce quantile-quant
## qqnorm(): produces a normal QQ plot of the variable
## qqline(): adds a reference line
## Use these two functions to create a QQ plot for sepal length. You will need to se
## colour as green, and its width as 2)
## qqnorm is used to creaate normal qq plot of sepal length variable. and i used pch
qqnorm(df$Sepal.Length, pch = 1, frame = FALSE)

## width 2 , color green
##qqline method adds a reference line to our qqnorm q plot drwaing
qqline(df$Sepal.Length, col = "green", lwd = 2)
```



In [ ]:
```
###                          --------  WEEK10 -------
```

In [ ]:
```
# installing necessary ggplot libraries.
```

In [45]:
```
install.packages("ggplot2")
```

```
  There is a binary version available but the source version is later:
        binary source needs_compilation
ggplot2  3.3.3  3.3.5              FALSE

installing the source package 'ggplot2'
```

In [44]:
```
## just seeing the overal table using the head() function to know more details about
head(mpg)
```

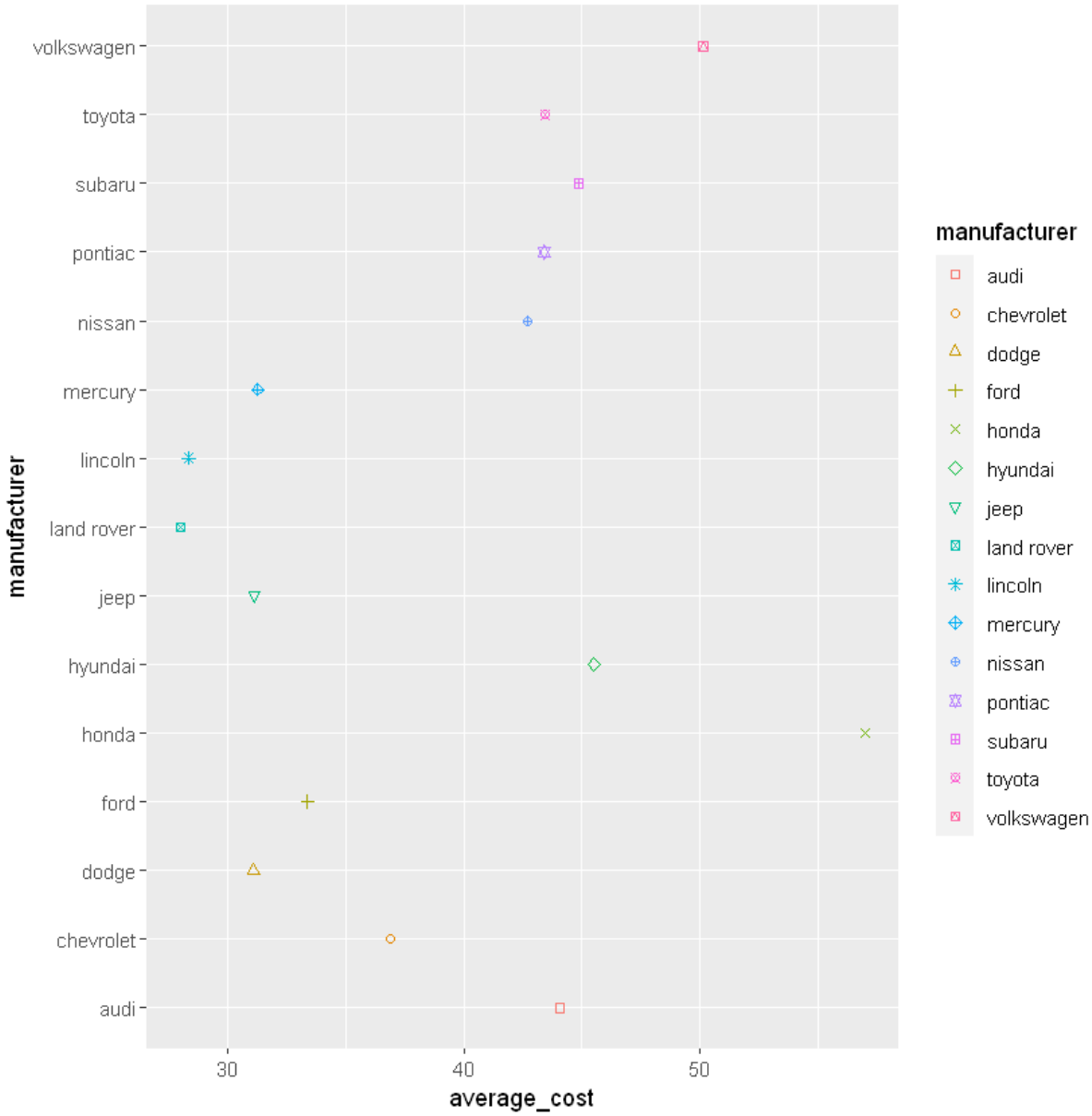
manufacturer	model	displ	year	cyl	trans	drv	cty	hwy	fl	class
audi	a4	1.8	1999	4	auto(l5)	f	18	29	p	compact
audi	a4	1.8	1999	4	manual(m5)	f	21	29	p	compact

manufacturer	model	displ	year	cyl	trans	drv	cty	hwy	fl	class
audi	a4	2.0	2008	4	manual(m6)	f	20	31	p	compact
audi	a4	2.0	2008	4	auto(av)	f	21	30	p	compact
audi	a4	2.8	1999	6	auto(l5)	f	16	26	p	compact
audi	a4	2.8	1999	6	manual(m5)	f	18	26	p	compact

In [46]:
```
##Q1. Plot and explain: Which vehicle brand (or manufacturer), offers the best mpg i
##both city and in the highway? (6 marks)
## I took the average of every brand for cty and hwy values and ploted them using gr
library(dplyr)
averagebrand<-mpg %>%
  group_by(manufacturer) %>%
  summarise(average_cost = mean(hwy+cty))
```

In [47]:
```
mpg_stat <- ggplot(averagebrand, aes(x = average_cost,y = manufacturer))
mpg_stat + geom_point(aes(shape = manufacturer,color=manufacturer)) +scale_shape_man
## I addde that section to be more readeble for colours and shapes manuaaly changed
  stat_smooth(method = "lm")
```

```
geom_smooth: se = TRUE, na.rm = FALSE, orientation = NA
stat_smooth: method = lm, formula = NULL, se = TRUE, n = 80, fullrange = FALSE, leve
l = 0.95, na.rm = FALSE, orientation = NA, method.args = list(), span = 0.75
position_identity
```



In [48]:

```
## just wanted to check th values for every brand
library(dplyr)
mpg %>%
  group_by(manufacturer) %>%
  summarise(average_cost = mean(hwy+cty))
```

manufacturer	average_cost
audi	44.05556
chevrolet	36.89474
dodge	31.08108
ford	33.36000
honda	57.00000
hyundai	45.50000
jeep	31.12500
land rover	28.00000
lincoln	28.33333

manufacturer	average_cost
mercury	31.25000
nissan	42.69231
pontiac	43.40000
subaru	44.85714
toyota	43.44118
volkswagen	50.14815

In [49]:
```
## Q.2 Plot and explain: Which type of car, regarding their displ range (size of eng
##has the lowest mpg in the city categorised by the vehicle type (e.g., compact, suv
##2seaters defined in the variable class)? Display the resulting plot categorised by
##vehicle type. (6 marks)
## Hint: facet_wrap() for the categorisation
```

In [50]:
```
library(dplyr)
## I assigned the average cost of city per engine size cost. Because question mentio
## engine size so I divided cty to displ(engine size). I also added displ to my data
my_car<-mpg %>%
  group_by(class,displ) %>%
  summarise(average_cost = mean(cty/displ))
my_car
```

```
`summarise()` has grouped output by 'class'. You can override using the `.groups` ar
gument.
```

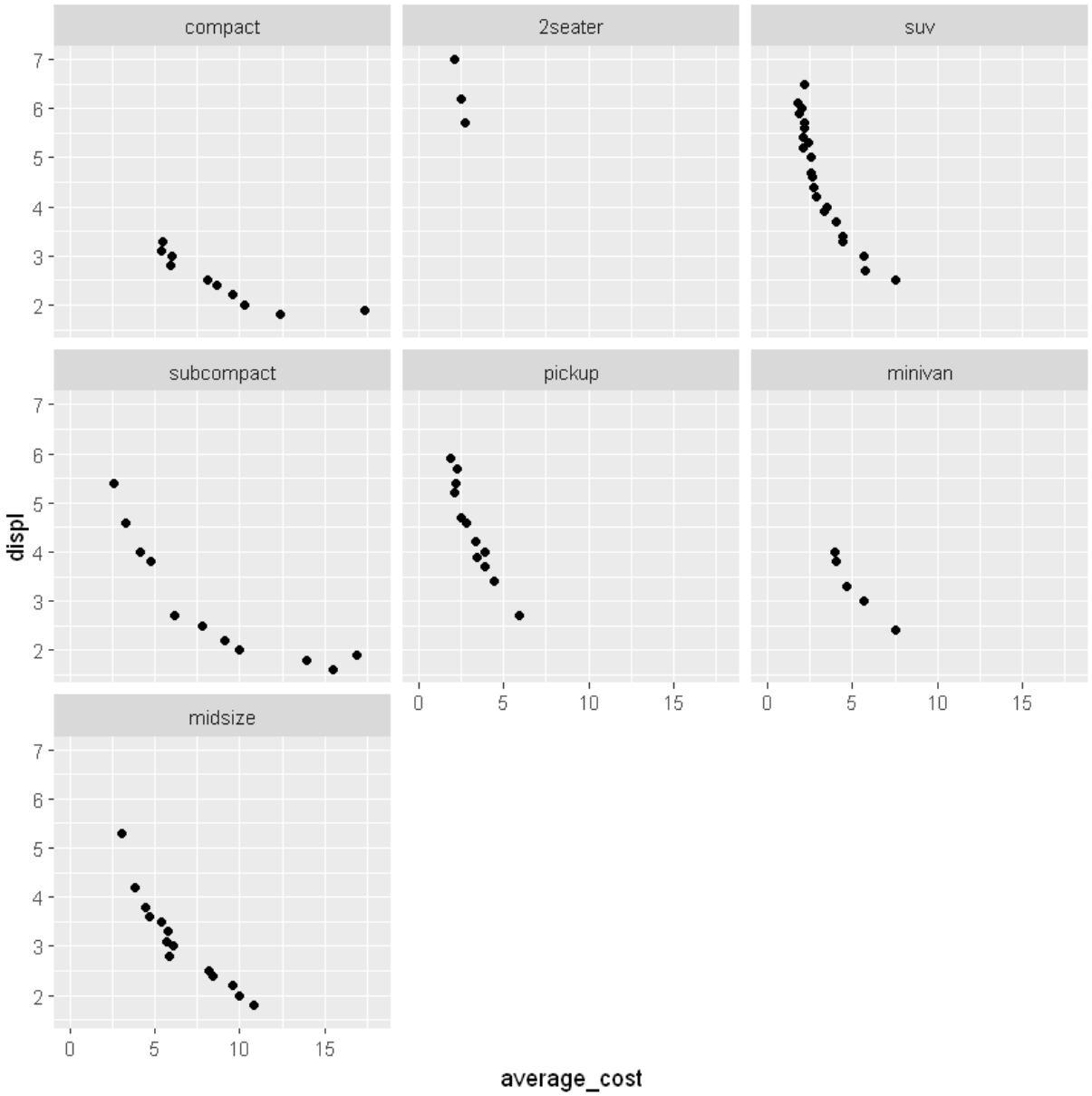
class	displ	average_cost
2seater	5.7	2.719298
2seater	6.2	2.500000
2seater	7.0	2.142857
compact	1.8	12.407407
compact	1.9	17.368421
compact	2.0	10.250000
compact	2.2	9.545455
compact	2.4	8.645833
compact	2.5	8.066667
compact	2.8	5.918367
compact	3.0	6.000000
compact	3.1	5.376344
compact	3.3	5.454545
midsize	1.8	10.833333
midsize	2.0	10.000000
midsize	2.2	9.545455
midsize	2.4	8.385417

class	displ	average_cost
midsize	2.5	8.200000
midsize	2.8	5.833333
midsize	3.0	6.083333
midsize	3.1	5.698925
midsize	3.3	5.757576
midsize	3.5	5.371429
midsize	3.6	4.722222
midsize	3.8	4.473684
midsize	4.2	3.809524
midsize	5.3	3.018868
minivan	2.4	7.500000
minivan	3.0	5.666667
minivan	3.3	4.666667
...	...	...
subcompact	2.0	9.928571
subcompact	2.2	9.090909
subcompact	2.5	7.800000
subcompact	2.7	6.172840
subcompact	3.8	4.736842
subcompact	4.0	4.125000
subcompact	4.6	3.260870
subcompact	5.4	2.592593
suv	2.5	7.533333
suv	2.7	5.740741
suv	3.0	5.666667
suv	3.3	4.393939
suv	3.4	4.411765
suv	3.7	4.054054
suv	3.9	3.333333
suv	4.0	3.475000
suv	4.2	2.857143
suv	4.4	2.727273
suv	4.6	2.608696
suv	4.7	2.579787
suv	5.0	2.600000
suv	5.2	2.115385

class	displ	average_cost
suv	5.3	2.415094
suv	5.4	2.111111
suv	5.6	2.142857
suv	5.7	2.210526
suv	5.9	1.864407
suv	6.0	2.000000
suv	6.1	1.803279
suv	6.5	2.153846

In [ ]:

In [51]:
```
my_car <- within(my_car, class <- factor(class, levels=c('compact', '2seater', 'suv'
                                                         'subcompact', 'pickup',
                                                         'minivan', 'midsize')))
#I customly put them order. I used facet_wrap() method to show each category for rel
ggplot(my_car, aes(average_cost,displ)) +xlim(0,18)+
  geom_point() +
  facet_wrap(vars(class))
```



In [52]:
```
##  Plot and explain: Which type of car, regarding their displ range (size of engine
## has the best mpg performance in both city and highway? Display the resulting plot
## categorised by the number of cylinders and the drive type (the type of drive trai
## where f = front-wheel drive, r = rear wheel drive, 4 = 4wd). You are a buyer who
## wants a high litre engine vehicle and drives mostly in the highway, which type of
## would you choose?
```

In [53]:
```
library(dplyr)
## I assigned the average cost of city per engine size cost. Because question mentio
## engine size so I divided cty to displ(engine size). I also added displ to my data
my_car2<-mpg %>%
  group_by(class,displ,cyl,drv) %>%
  summarise(average_cost = mean(mean(cty+hwy)/displ))
my_car2
```

```
`summarise()` has grouped output by 'class', 'displ', 'cyl'. You can override using
the `.groups` argument.
```

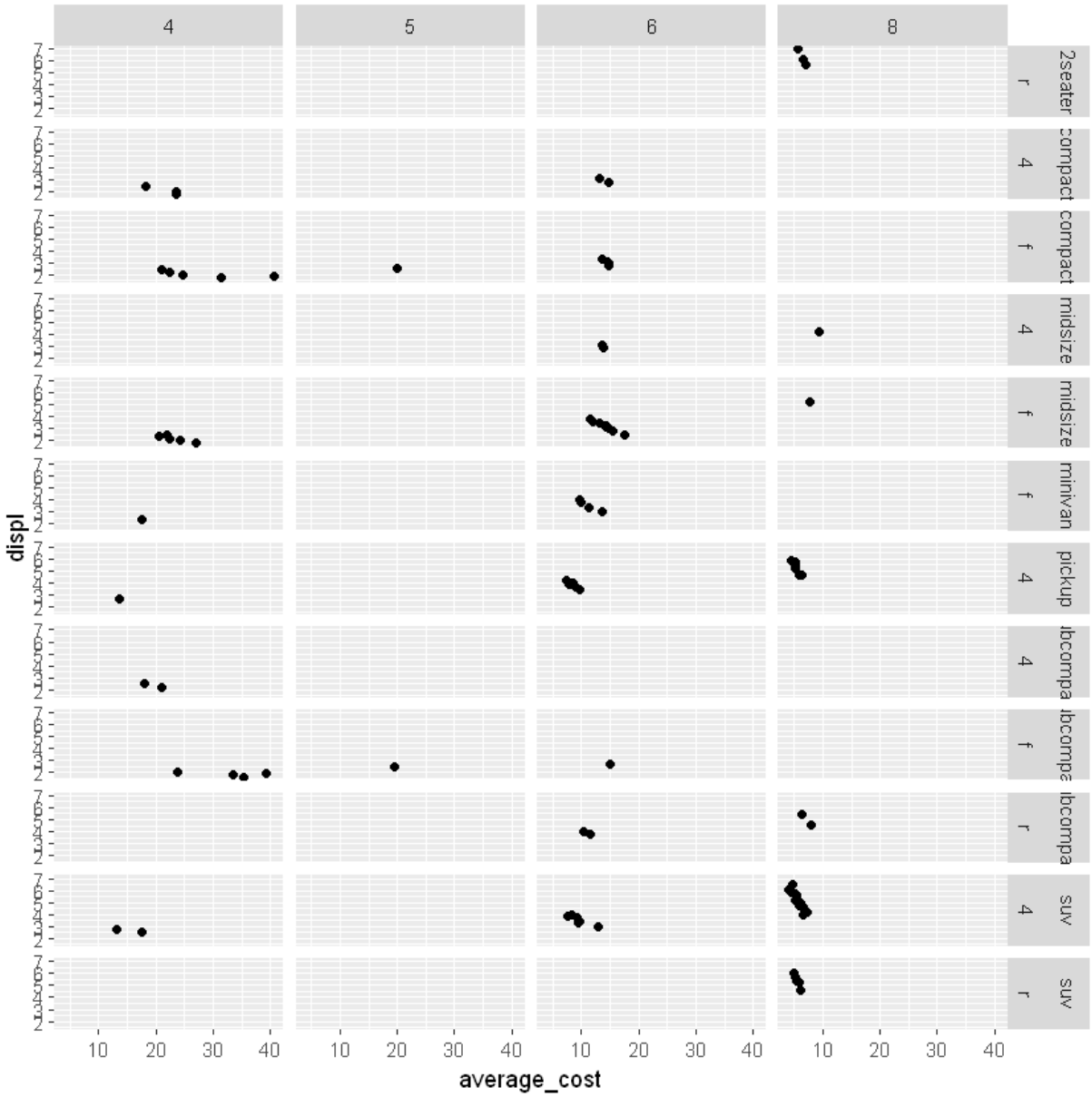
class	displ	cyl	drv	average_cost
2seater	5.7	8	r	7.017544
2seater	6.2	8	r	6.612903

class	displ	cyl	drv	average_cost
2seater	7.0	8	r	5.571429
compact	1.8	4	4	23.611111
compact	1.8	4	f	31.349206
compact	1.9	4	f	40.526316
compact	2.0	4	4	23.500000
compact	2.0	4	f	24.700000
compact	2.2	4	f	22.272727
compact	2.4	4	f	20.937500
compact	2.5	4	4	18.300000
compact	2.5	5	f	20.000000
compact	2.8	6	4	14.642857
compact	2.8	6	f	14.785714
compact	3.0	6	f	14.666667
compact	3.1	6	4	13.225806
compact	3.1	6	f	14.516129
compact	3.3	6	f	13.636364
midsize	1.8	4	f	26.944444
midsize	2.0	4	f	24.250000
midsize	2.2	4	f	22.272727
midsize	2.4	4	f	20.520833
midsize	2.5	4	f	21.800000
midsize	2.5	6	f	17.600000
midsize	2.8	6	4	13.928571
midsize	2.8	6	f	15.357143
midsize	3.0	6	f	14.666667
midsize	3.1	6	4	13.548387
midsize	3.1	6	f	14.193548
midsize	3.3	6	f	14.242424
...	...	...	...	...
subcompact	3.8	6	r	11.447368
subcompact	4.0	6	r	10.375000
subcompact	4.6	8	r	8.043478
subcompact	5.4	8	r	6.296296
suv	2.5	4	4	17.533333
suv	2.7	4	4	13.148148
suv	3.0	6	4	13.000000

class	displ	cyl	drv	average_cost
suv	3.3	6	4	9.545455
suv	3.4	6	4	9.705882
suv	3.7	6	4	9.189189
suv	3.9	6	4	7.692308
suv	4.0	6	4	8.222222
suv	4.0	8	4	6.500000
suv	4.2	8	4	7.142857
suv	4.4	8	4	6.818182
suv	4.6	8	4	6.521739
suv	4.6	8	r	6.086957
suv	4.7	8	4	5.930851
suv	5.0	8	4	6.000000
suv	5.2	8	4	5.192308
suv	5.3	8	4	5.471698
suv	5.3	8	r	5.911950
suv	5.4	8	r	5.296296
suv	5.6	8	4	5.357143
suv	5.7	8	4	5.219298
suv	5.7	8	r	5.263158
suv	5.9	8	4	4.406780
suv	6.0	8	r	4.833333
suv	6.1	8	4	4.098361
suv	6.5	8	4	4.769231

In [54]:

```
## I added two more dimensions to my facegrit to according to needs in question whic
## iw till be more like a 3d table and average cost values for city and highway also
## I plotted x axis as average_cost and y axis as displ(engine size) but I also cate
ggplot(data = my_car2) +
  geom_point(mapping = aes(x = average_cost, y = displ)) +
  facet_grid(class ~ drv ~ cyl)
```



In [55]:
```
## According to table for a high litre car choice- it looks logical to buy "8 cyland
```

In [ ]:
```

```