(https://github.com/dd-consulting)

One-Stop Analytics: R

Case Study of Autism Spectrum Disorder (ASD) with R

ABOUT 1 IN 59 CHILDREN

WERE IDENTIFIED WITH AUTISM SPECTRUM DISORDER
AMONG A 2014 SAMPLE OF 8 YEAR OLDS FROM 11 US COMMUNITIES
IN CDC'S ADDM NETWORK

[United States]

Centers for Disease Control and Prevention (CDC) - Autism Spectrum Disorder (ASD)

Autism spectrum disorder (ASD) is a developmental disability that can cause significant social, communication and behavioral challenges. CDC is committed to continuing to provide essential data on ASD, search for factors that put children at risk for ASD and possible causes, and develop resources that help identify children with ASD as early as possible.

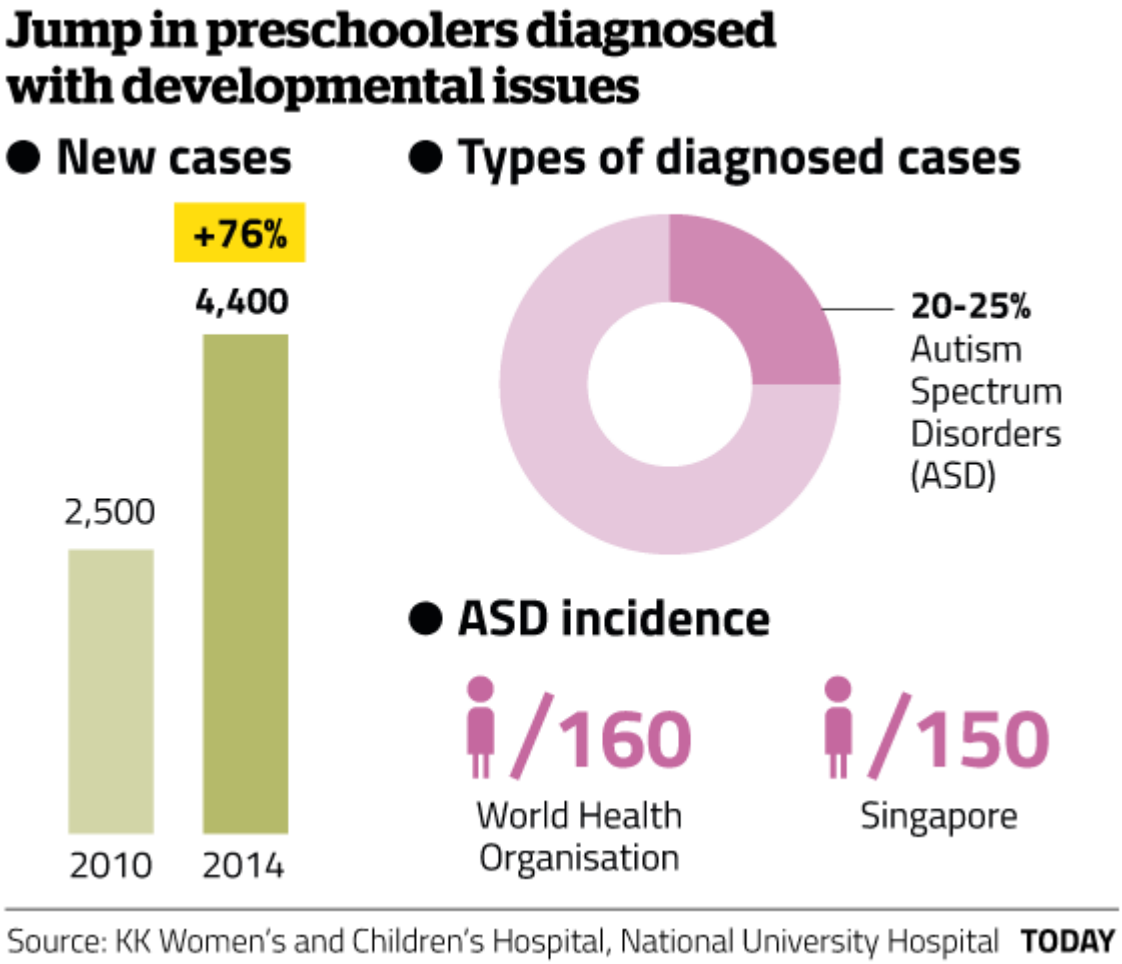
https://www.cdc.gov/ncbddd/autism/data/index.html (https://www.cdc.gov/ncbddd/autism/data/index.html)

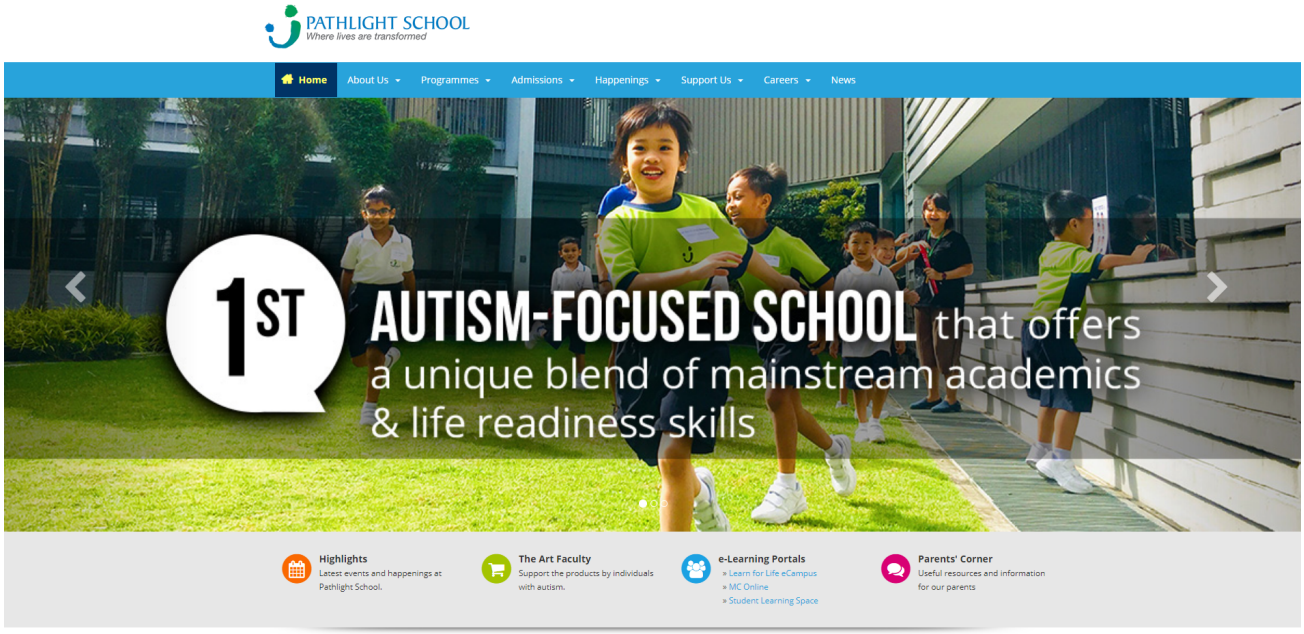
[Singapore]

TODAY Online - More preschoolers diagnosed with developmental issues

Doctors cited better awareness among parents and preschool teachers, leading to early referrals for diagnosis.

https://www.gov.sg/news/content/today-online-more-preschoolers-diagnosed-with-developmental-issues (https://www.gov.sg/news/content/today-online-more-preschoolers-diagnosed-with-developmental-issues)





https://www.pathlight.org.sg/ (https://www.pathlight.org.sg/)

()

Workshop Objective:

Use R to analyze Autism Spectrum Disorder (ASD) data from CDC USA.

https://www.cdc.gov/ncbddd/autism/data/index.html (https://www.cdc.gov/ncbddd/autism/data/index.html)

- **R Fundamentals**
- **Data Summarization**
- **Data Visualisation (Base Graphic)**
- **Appendices**

()

R Fundamentals

R Fundamentals - Get & Set working directory

Obtain current R working directory

In [1]:
```
getwd()
```

'/media/sf_vm_shared_folder/git/DDC/DDC-ASD/model_R'

Set new R working directory

In [2]:
```
# setwd("/media/sf_vm_shared_folder/git/DDC/DDC-ASD/model_R")
# setwd('~/Desktop/admin-desktop/vm_shared_folder/git/DDC-ASD/model_R')
getwd()
```

'/media/sf_vm_shared_folder/git/DDC/DDC-ASD/model_R'

Read in CSV data, storing as R dataframe

In [3]:
```
# Dataset: US. National Level Children ASD Prevalence
ASD_National <- read.csv("../dataset/ADV_ASD_National.csv", stringsAsFactors =
```

In [4]:
```
# Dataset: US. State Level Children ASD Prevalence
ASD_State    <- read.csv("../dataset/ADV_ASD_State.csv", stringsAsFactors = FA
```

Look at first/last few rows of data

In [5]:
```
head(ASD_National)
```

Source	Year	Prevalence	Upper.CI	Lower.CI	Prevalence_dup	Source_Full1	Source_Full2	Male.Preval
addm	2000	6.7	7.0	6.3	6.7	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	Nc
addm	2002	6.6	6.8	6.3	6.6	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	1
addm	2004	8.0	8.4	7.6	8.0	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	1
addm	2006	9.0	9.3	8.6	9.0	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	1
addm	2008	11.3	11.7	11.0	11.3	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	1
addm	2010	14.7	15.1	14.3	14.7	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	2

In [6]:
```
tail(ASD_State)
```

	State	Denominator	Prevalence	Lower.CI	Upper.CI	Year	Source	Source_Full1	State_Full1	Stat
1687	UT	596257	8.7	8.5	9.0	2016	sped	Special Education Child Count	Utah	U
1688	VT	74108	12.1	11.3	12.9	2016	sped	Special Education Child Count	Vermont	VT-V
1689	VA	1162945	14.2	14.0	14.4	2016	sped	Special Education Child Count	Virginia	VA-
1690	WA	1006676	11.2	11.0	11.4	2016	sped	Special Education Child Count	Washington	Was
1691	WV	239037	8.6	8.3	9.0	2016	sped	Special Education Child Count	West Virginia	W
1692	WY	85922	9.3	8.7	10.0	2016	sped	Special Education Child Count	Wyoming	W

Obtain number of rows and number of columns/features/variables

In [7]:
```
dim(ASD_National)
```

```
42  26
```

In [8]:
```
dim(ASD_State)
```

```
1692  49
```

Obtain overview (data structure/types)

In [9]:
```
str(ASD_National)
```

```
'data.frame':   42 obs. of  26 variables:
 $ Source                                : chr  "addm" "addm" "addm" "addm"
...
 $ Year                                  : int  2000 2002 2004 2006 2008 2010
2012 2014 2004 2008 ...
 $ Prevalence                            : num  6.7 6.6 8 9 11.3 14.7 14.8 1
6.8 9.5 16.2 ...
 $ Upper.CI                              : num  7 6.8 8.4 9.3 11.7 15.1 15.2
17.3 12 18.1 ...
 $ Lower.CI                              : num  6.3 6.3 7.6 8.6 11 14.3 14.4
16.4 7.4 14.5 ...
 $ Prevalence_dup                        : num  6.7 6.6 8 9 11.3 14.7 14.8 1
6.8 9.5 16.2 ...
 $ Source_Full1                          : chr  "Autism & Developmental Disab
ilities Monitoring Network" "Autism & Developmental Disabilities Monitoring
Network" "Autism & Developmental Disabilities Monitoring Network" "Autism &
Developmental Disabilities Monitoring Network" ...
 $ Source_Full2                          : chr  "addm-Autism & Developmental
Disabilities Monitoring Network" "addm-Autism & Developmental Disabilities
```

In [10]:
```
str(ASD_State)
```

```
'data.frame':   1692 obs. of  49 variables:
 $ State                                 : chr  "AZ" "GA" "MD" "NJ" ...
 $ Denominator                           : int  45322 43593 21532 29714 245
35 23065 35472 45113 36472 11020 ...
 $ Prevalence                            : num  6.5 6.5 5.5 9.9 6.3 4.5 3.3
6.2 6.9 5.9 ...
 $ Lower.CI                              : num  5.8 5.8 4.6 8.9 5.4 3.7 2.7
5.5 6.1 4.6 ...
 $ Upper.CI                              : num  7.3 7.3 6.6 11.1 7.4 5.5 3.
9 7 7.8 7.5 ...
 $ Year                                  : int  2000 2000 2000 2000 2000 20
00 2002 2002 2002 2002 ...
 $ Source                                : chr  "addm" "addm" "addm" "addm"
...
 $ Source_Full1                          : chr  "Autism & Developmental Dis
abilities Monitoring Network" "Autism & Developmental Disabilities Monitori
ng Network" "Autism & Developmental Disabilities Monitoring Network" "Autis
m & Developmental Disabilities Monitoring Network" ...
 $ State_Full1                           : chr  "Arizona" "Georgia" "Maryla
```

Obtain name of columns

In [11]:
```
names(ASD_National)
```

'Source' 'Year' 'Prevalence' 'Upper.CI' 'Lower.CI' 'Prevalence_dup' 'Source_Full1'
'Source_Full2' 'Male.Prevalence' 'Male.Lower.CI' 'Male.Upper.CI' 'Female.Prevalence'
'Female.Lower.CI' 'Female.Upper.CI' 'Non.hispanic.white.Prevalence' 'Non.hispanic.white.Lower.CI'
'Non.hispanic.white.Upper.CI' 'Non.hispanic.black.Prevalence' 'Non.hispanic.black.Lower.CI'
'Non.hispanic.black.Upper.CI' 'Hispanic.Prevalence' 'Hispanic.Lower.CI' 'Hispanic.Upper.CI'
'Asian.or.Pacific.Islander.Prevalence' 'Asian.or.Pacific.Islander.Lower.CI'
'Asian.or.Pacific.Islander.Upper.CI'

In [12]:
```
names(ASD_State)
```

'State' 'Denominator' 'Prevalence' 'Lower.CI' 'Upper.CI' 'Year' 'Source' 'Source_Full1'
'State_Full1' 'State_Full2' 'Numerator_ASD' 'Numerator_NonASD' 'Proportion' 'X95_Z_CI'
'Z_Lower.CI' 'Z_Upper.CI' 'Z_Lower.CI_ABSerror' 'Z_Upper.CI_ABSerror' 'Chi_Wilson_P'
'X95_Chi_Wilson_CI' 'Chi_Wilson_Lower.CI' 'Chi_Wilson_Upper.CI'
'Chi_Wilson_Lower.CI_ABSerror' 'Chi_Wilson_Upper.CI_ABSerror'
'Chi_Wilson_Corrected_w_minus.CI' 'Chi_Wilson_Corrected_w_plus.CI'
'Chi_Wilson_Corrected_Lower.CI' 'Chi_Wilson_Corrected_Upper.CI'
'Chi_Wilson_Corrected_Lower.CI_ABSerror' 'Chi_Wilson_Corrected_Upper.CI_ABSerror'
'Male.Prevalence' 'Male.Lower.CI' 'Male.Upper.CI' 'Female.Prevalence' 'Female.Lower.CI'
'Female.Upper.CI' 'Non.hispanic.white.Prevalence' 'Non.hispanic.white.Lower.CI'
'Non.hispanic.white.Upper.CI' 'Non.hispanic.black.Prevalence' 'Non.hispanic.black.Lower.CI'
'Non.hispanic.black.Upper.CI' 'Hispanic.Prevalence' 'Hispanic.Lower.CI' 'Hispanic.Upper.CI'
'Asian.or.Pacific.Islander.Prevalence' 'Asian.or.Pacific.Islander.Lower.CI'
'Asian.or.Pacific.Islander.Upper.CI' 'State_Region'

Display column name with its index number

In [13]:
```
cbind(names(ASD_National), c(1:length(names(ASD_National))))
```

Source	1
Year	2
Prevalence	3
Upper.CI	4
Lower.CI	5
Prevalence_dup	6
Source_Full1	7
Source_Full2	8
Male.Prevalence	9
Male.Lower.CI	10
Male.Upper.CI	11
Female.Prevalence	12
Female.Lower.CI	13

Look at data structure/schema (Selected columns)

In [14]:
```
str(ASD_National[, c(1:8, 24, 25, 26)])
```

```
'data.frame':   42 obs. of  11 variables:
 $ Source                                : chr  "addm" "addm" "addm" "addm" ...
 $ Year                                  : int  2000 2002 2004 2006 2008 2010 2
012 2014 2004 2008 ...
 $ Prevalence                            : num  6.7 6.6 8 9 11.3 14.7 14.8 16.8
9.5 16.2 ...
 $ Upper.CI                              : num  7 6.8 8.4 9.3 11.7 15.1 15.2 1
7.3 12 18.1 ...
 $ Lower.CI                              : num  6.3 6.3 7.6 8.6 11 14.3 14.4 1
6.4 7.4 14.5 ...
 $ Prevalence_dup                        : num  6.7 6.6 8 9 11.3 14.7 14.8 16.8
9.5 16.2 ...
 $ Source_Full1                          : chr  "Autism & Developmental Disabil
ities Monitoring Network" "Autism & Developmental Disabilities Monitoring Net
work" "Autism & Developmental Disabilities Monitoring Network" "Autism & Deve
lopmental Disabilities Monitoring Network" ...
 $ Source_Full2                          : chr  "addm-Autism & Developmental Di
sabilities Monitoring Network" "addm-Autism & Developmental Disabilities Moni
toring Network" "addm-Autism & Developmental Disabilities Monitoring Network"
"addm-Autism & Developmental Disabilities Monitoring Network" ...
 $ Asian.or.Pacific.Islander.Prevalence: chr  "No data" "No data" "No data"
"No data" ...
 $ Asian.or.Pacific.Islander.Lower.CI  : chr  "No data" "No data" "No data"
"No data" ...
 $ Asian.or.Pacific.Islander.Upper.CI  : chr  "No data" "No data" "No data"
"No data" ...
```

Quiz:

Obtain feature/column names and column index of dataframe: ASD_State

In [15]:
```
# Write your code below and press Shift+Enter to execute
```

Double-click **here** for the solution.

R Fundamentals - Work with dataframe

Access column 1 as a named list:

In [16]:
```
# use column index:
ASD_National[1]
```

Source
addm
addm
addm
addm
addm
addm
addm
addm
nsch
nsch
nsch
nsch

In [17]:
```
typeof(ASD_National[1])
```

'list'

In [18]:
```
ASD_National[1]$Source
```

'addm' 'addm' 'addm' 'addm' 'addm' 'addm' 'addm' 'addm' 'nsch' 'nsch' 'nsch' 'nsch'
'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped'
'sped' 'sped' 'sped' 'sped' 'medi' 'medi' 'medi' 'medi' 'medi' 'medi' 'medi' 'medi' 'medi'
'medi' 'medi' 'medi' 'medi'

In [19]:
```
typeof(ASD_National[1]$Source)
```

'character'

In [20]:
```
# use column name:
ASD_National["Source"]
```

Source
addm
addm
addm
addm
addm
addm
addm
addm
nsch
nsch
nsch
nsch

In [21]:
```
ASD_National['Source']$Source
```

'addm' 'addm' 'addm' 'addm' 'addm' 'addm' 'addm' 'addm' 'nsch' 'nsch' 'nsch' 'nsch'
'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped'
'sped' 'sped' 'sped' 'sped' 'medi' 'medi' 'medi' 'medi' 'medi' 'medi' 'medi' 'medi' 'medi'
'medi' 'medi' 'medi' 'medi'

Access column 1 as a set of string/chr:

In [22]:
```
ASD_National[, 1]
```

'addm' 'addm' 'addm' 'addm' 'addm' 'addm' 'addm' 'addm' 'nsch' 'nsch' 'nsch' 'nsch'
'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped'
'sped' 'sped' 'sped' 'sped' 'medi' 'medi' 'medi' 'medi' 'medi' 'medi' 'medi' 'medi' 'medi'
'medi' 'medi' 'medi' 'medi'

In [23]:
```
# or
ASD_National[, "Source"]
```

'addm' 'addm' 'addm' 'addm' 'addm' 'addm' 'addm' 'addm' 'nsch' 'nsch' 'nsch' 'nsch'
'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped'
'sped' 'sped' 'sped' 'sped' 'medi' 'medi' 'medi' 'medi' 'medi' 'medi' 'medi' 'medi' 'medi'
'medi' 'medi' 'medi' 'medi'

In [24]:
```
# or
ASD_National$Source
```

'addm' 'addm' 'addm' 'addm' 'addm' 'addm' 'addm' 'addm' 'nsch' 'nsch' 'nsch' 'nsch'
'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped'
'sped' 'sped' 'sped' 'sped' 'medi' 'medi' 'medi' 'medi' 'medi' 'medi' 'medi' 'medi' 'medi'
'medi' 'medi' 'medi' 'medi'

In [25]:
```
typeof(ASD_National$Source)
```

'character'

Count number of elements in a object:

In [26]:
```
length(ASD_National) # number of features/columns
```

26

In [27]:
```
length(ASD_National[1, ]) # number of elements(columns) in row 1
```

26

In [28]:
```
length(ASD_National[, 1]) # number of elements(rows) in column 1
```

42

In [29]:
```
length(ASD_National[, "Source"]) # same as above
```

42

In [30]:
```
length(ASD_National$Source) # number of elements in chr list
```

42

Access elements from dataframe

In [31]:
```
# using column index
ASD_National[1][1, ]
```

'addm'

In [32]:
```
ASD_National[1][11, ]
```

'nsch'

In [33]:
```
ASD_National[1][11:20, ]
```

'nsch' 'nsch' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped'

In [34]:
```
# using column name
ASD_National["Source"][1, ]
```

'addm'

In [35]:
```
ASD_National["Source"][11, ]
```

'nsch'

In [36]:
```
ASD_National["Source"][11:20, ]
```

'nsch' 'nsch' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped'

Access elements from dataframe

In [37]:
```
# using column index
ASD_National[, 1][1]
```

'addm'

In [38]:
```
ASD_National[, 1][11]
```

'nsch'

In [39]:
```
ASD_National[, 1][11:20]
```

'nsch' 'nsch' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped'

In [40]:
```
# using column name
ASD_National[, "Source"][1]
```

'addm'

In [41]:
```
# using column name
ASD_National[, "Source"][11]
```

'nsch'

In [42]:
```
# using column name
ASD_National[, "Source"][11:20]
```

'nsch' 'nsch' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped'

In [43]:
```
# using $ operator
ASD_National$Source[1]
```

'addm'

In [44]:
```
ASD_National$Source[11]
```

'nsch'

In [45]:
```
ASD_National$Source[11:20]
```

'nsch' 'nsch' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped' 'sped'

Access elements of different column:

In [46]:
```
cbind(names(ASD_National), c(1:length(names(ASD_National))))
```

Source	1
Year	2
Prevalence	3
Upper.CI	4
Lower.CI	5
Prevalence_dup	6
Source_Full1	7
Source_Full2	8
Male.Prevalence	9
Male.Lower.CI	10
Male.Upper.CI	11
Female.Prevalence	12
Female.Lower.CI	13

In [47]:
```
ASD_National[1, 1] # row 1, column 1: "Source"
```

'addm'

In [48]:
```
ASD_National[10, 1] # row 10, column 1: "Source"
```

'nsch'

In [49]:
```
ASD_National[1, 3] # row 1, column 3: "Prevalence"
```

6.7

In [50]:
```
ASD_National[10, 3] # row 10, column 3: "Prevalence"
```

16.2

In [51]:
```
ASD_National[1:10, 1:3] # row 1 to 10 from column 1 to 3
```

Source	Year	Prevalence
addm	2000	6.7
addm	2002	6.6
addm	2004	8.0
addm	2006	9.0
addm	2008	11.3
addm	2010	14.7
addm	2012	14.8
addm	2014	16.8
nsch	2004	9.5
nsch	2008	16.2

In [52]:
```
# or using columns names
ASD_National[1:10, c('Source', 'Year', 'Prevalence')]
```

Source	Year	Prevalence
addm	2000	6.7
addm	2002	6.6
addm	2004	8.0
addm	2006	9.0
addm	2008	11.3
addm	2010	14.7
addm	2012	14.8
addm	2014	16.8
nsch	2004	9.5
nsch	2008	16.2

In [53]:
```
ASD_National[c(1:10, 20, 30:35), c(1:3, 9, 12)] # row 1 to 10, 20, and 20 to 2
```

	Source	Year	Prevalence	Male.Prevalence	Female.Prevalence
1	addm	2000	6.7	No data	No data
2	addm	2002	6.6	11.5	2.7
3	addm	2004	8.0	12.9	2.9
4	addm	2006	9.0	14.5	3.2
5	addm	2008	11.3	18.4	4
6	addm	2010	14.7	23.7	5.3
7	addm	2012	14.8	23.4	5.2
8	addm	2014	16.8	26.6	6.6
9	nsch	2004	9.5		
10	nsch	2008	16.2		
20	sped	2007	5.4		
30	medi	2000	2.3		
31	medi	2001	2.6		
32	medi	2002	2.8		
33	medi	2003	3.0		
34	medi	2004	3.5		
35	medi	2005	3.9		

[Tips] We notice missing data from above.

R Fundamentals - Process missing data

Count missing values in dataframe:

In [54]:
```
sum(is.na(ASD_National)) # No missing data recognised by R (NA)
```

0

In [55]:
```
sum(is.na(ASD_State)) # Some missing data recognised by R (NA)
```

14454

Empty string, "No data" are not considered as missing value by R, thus we need to handle them manually.

In [56]:
```
# Define several offending strings
na_strings <- c("", "No data", "NA", "N A", "N / A", "N/A", "N/ A", "Not Avail
```

In [57]:
```
# Load required function from packages:
if(!require(naniar)){install.packages("naniar")}
library(naniar)
if(!require(dplyr)){install.packages("dplyr")}
library(dplyr)
```

```
Loading required package: naniar
Registered S3 methods overwritten by 'ggplot2':
  method         from
  [.quosures     rlang
  c.quosures     rlang
  print.quosures rlang
Loading required package: dplyr

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 [58]:
```
# Uncomment below to show help
# ?replace_with_na_all # Documentation
```

Replace these defined missing/offending values to R's internal NA

In [59]:
```
# "~.x" is a reserved keyword of this function:
ASD_National = replace_with_na_all(ASD_National, condition = ~.x %in% na_strin
```

In [60]:
```
# Count missing values (R's internal NA) in dataframe:
sum(is.na(ASD_National))
```

650

R Fundamentals - Process invalid characters

Remove invalid unicode char/string: \x92

In [61]:
```
ASD_National$Source_Full1[ASD_National$Source_Full1 == "National Survey of Chi
"National Survey of Children's Health"
```

In [62]:
```
ASD_National$Source_Full2[ASD_National$Source_Full2 == "nsch-National Survey o
"nsch-National Survey of Children's Health"
```

R Fundamentals - Delete/Drop dataframe variable

Delete/Drop duplicate variable: Prevalence_dup

In [63]:
```
drop <- c("Prevalence_dup", "Dummy Variable Name")
```

In [64]:
```
ASD_National = ASD_National[, !(names(ASD_National) %in% drop)] # Recall Dataf
```

R Fundamentals - Create/Add dataframe variable

Create one new variable: Source_UC by converting to uppercase letters

In [65]:
```
ASD_National$Source_UC <- paste(toupper(ASD_National$Source))
```

Create one new variable: Source_Full3 by combining Source and Source_Full1

In [66]:
```
ASD_National$Source_Full3 <- paste(toupper(ASD_National$Source), ASD_National$
```

Create one new ordinal categorical variable: Prevalence_Rank2 ("Low", "High") by binning Prevalence

In [67]:
```
# Recode Risk into category from Prevalence

# Low [0, 5)
# High [5, +oo)

ASD_National$Prevalence_Risk2[ASD_National$Prevalence < 5] = "Low"
ASD_National$Prevalence_Risk2[ASD_National$Prevalence >= 5 ] = "High"
#
head(ASD_National)
```

```
Warning message:
“Unknown or uninitialised column: 'Prevalence_Risk2'.”
```

Source	Year	Prevalence	Upper.CI	Lower.CI	Source_Full1	Source_Full2	Male.Prevalence	Male.Lower
addm	2000	6.7	7.0	6.3	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	NA	N
addm	2002	6.6	6.8	6.3	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	11.5	N
addm	2004	8.0	8.4	7.6	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	12.9	1
addm	2006	9.0	9.3	8.6	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	14.5	1
addm	2008	11.3	11.7	11.0	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	18.4	1
addm	2010	14.7	15.1	14.3	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	23.7	2

Create one new ordinal categorical variable: Prevalence_Rank4 ("Low", "Medium", "High", "Very High") by binning Prevalence

In [68]:
```
# Recode Risk into category from Prevalence

# Low [0, 5)
# Medium [5, 10)
# High [10, 20)
# Very High [20, +oo)

ASD_National$Prevalence_Risk4 = "Very High"
ASD_National$Prevalence_Risk4[ASD_National$Prevalence < 20 ] = "High"
ASD_National$Prevalence_Risk4[ASD_National$Prevalence < 10 ] = "Medium"
ASD_National$Prevalence_Risk4[ASD_National$Prevalence < 5] = "Low"
#
head(ASD_National)
```

Source	Year	Prevalence	Upper.CI	Lower.CI	Source_Full1	Source_Full2	Male.Prevalence	Male.Lower
addm	2000	6.7	7.0	6.3	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	NA	N
addm	2002	6.6	6.8	6.3	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	11.5	N
addm	2004	8.0	8.4	7.6	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	12.9	1
addm	2006	9.0	9.3	8.6	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	14.5	1
addm	2008	11.3	11.7	11.0	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	18.4	1
addm	2010	14.7	15.1	14.3	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	23.7	2

R Fundamentals - Convert to correct data types

Review data structure and variable names:

In [69]:
```
str(ASD_National)
cbind(names(ASD_National), c(1:length(names(ASD_National))))
```

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       42 obs. of  29 variables:
 $ Source                                : chr  "addm" "addm" "addm" "addm" ...
 $ Year                                  : int  2000 2002 2004 2006 2008 2010 2
012 2014 2004 2008 ...
 $ Prevalence                            : num  6.7 6.6 8 9 11.3 14.7 14.8 16.8
9.5 16.2 ...
 $ Upper.CI                              : num  7 6.8 8.4 9.3 11.7 15.1 15.2 1
7.3 12 18.1 ...
 $ Lower.CI                              : num  6.3 6.3 7.6 8.6 11 14.3 14.4 1
6.4 7.4 14.5 ...
 $ Source_Full1                          : chr  "Autism & Developmental Disabil
ities Monitoring Network" "Autism & Developmental Disabilities Monitoring Net
work" "Autism & Developmental Disabilities Monitoring Network" "Autism & Deve
lopmental Disabilities Monitoring Network" ...
 $ Source_Full2                          : chr  "addm-Autism & Developmental Di
sabilities Monitoring Network" "addm-Autism & Developmental Disabilities Moni
toring Network" "addm-Autism & Developmental Disabilities Monitoring Network"
"addm-Autism & Developmental Disabilities Monitoring Network" ...
 $ Male.Prevalence                       : chr  NA "11.5" "12.9" "14.5" ...
 $ Male.Lower.CI                         : chr  NA NA "12.2" "13.9" ...
 $ Male.Upper.CI                         : chr  NA NA "13.7" "15.1" ...
 $ Female.Prevalence                     : chr  NA "2.7" "2.9" "3.2" ...
 $ Female.Lower.CI                       : chr  NA NA "2.6" "2.9" ...
 $ Female.Upper.CI                       : chr  NA NA "3.3" "3.5" ...
 $ Non.hispanic.white.Prevalence         : chr  NA "7.7" "9.7" "9.9" ...
 $ Non.hispanic.white.Lower.CI           : chr  NA NA "9.1" "9.4" ...
 $ Non.hispanic.white.Upper.CI           : chr  NA NA "10.4" "10.4" ...
 $ Non.hispanic.black.Prevalence         : chr  NA "6.5" "6.9" "7.2" ...
 $ Non.hispanic.black.Lower.CI           : chr  NA NA "6.2" "6.6" ...
 $ Non.hispanic.black.Upper.CI           : chr  NA NA "7.6" "7.8" ...
 $ Hispanic.Prevalence                   : chr  NA NA "6.2" "5.9" ...
 $ Hispanic.Lower.CI                     : chr  NA NA "5" "5.3" ...
 $ Hispanic.Upper.CI                     : chr  NA NA "7.5" "6.6" ...
 $ Asian.or.Pacific.Islander.Prevalence: chr  NA NA NA NA ...
 $ Asian.or.Pacific.Islander.Lower.CI  : chr  NA NA NA NA ...
 $ Asian.or.Pacific.Islander.Upper.CI  : chr  NA NA NA NA ...
 $ Source_UC                             : chr  "ADDM" "ADDM" "ADDM" "ADDM" ...
 $ Source_Full3                          : chr  "ADDM Autism & Developmental Di
sabilities Monitoring Network" "ADDM Autism & Developmental Disabilities Moni
toring Network" "ADDM Autism & Developmental Disabilities Monitoring Network"
"ADDM Autism & Developmental Disabilities Monitoring Network" ...
 $ Prevalence_Risk2                      : chr  "High" "High" "High" "High" ...
 $ Prevalence_Risk4                      : chr  "Medium" "Medium" "Medium" "Med
ium" ...
```

Source	1
Year	2
Prevalence	3
Upper.CI	4
Lower.CI	5
Source_Full1	6
Source_Full2	7
Male.Prevalence	8
Male.Lower.CI	9
Male.Upper.CI	10
Female.Prevalence	11

Female.Lower.CI	12
Female.Upper.CI	13
Non.hispanic.white.Prevalence	14
Non.hispanic.white.Lower.CI	15
Non.hispanic.white.Upper.CI	16
Non.hispanic.black.Prevalence	17
Non.hispanic.black.Lower.CI	18
Non.hispanic.black.Upper.CI	19
Hispanic.Prevalence	20
Hispanic.Lower.CI	21
Hispanic.Upper.CI	22
Asian.or.Pacific.Islander.Prevalence	23
Asian.or.Pacific.Islander.Lower.CI	24
Asian.or.Pacific.Islander.Upper.CI	25
Source_UC	26
Source_Full3	27
Prevalence_Risk2	28
Prevalence_Risk4	29

Convert Prevalence and CIs from categorical/chr to numeric, column 8 to 25

In [70]:
```
ix <- 8:25 # define an index
# apply()
ASD_National[ix] <- apply(ASD_National[ix], 2, as.numeric) # "2" meand column-
```

In [71]:
```
# Uncomment below to show help
# ?apply # Documentation
```

In [72]:
```
# or lapply()
ASD_National[ix] <- lapply(ASD_National[ix], as.numeric) # column-wise
```

In [73]:
```
# Uncomment below to show help
# ?lapply # Documentation
```

Convert Source from categorical/chr to categorical/factor

In [74]:
```
ix <- c(1, 6, 7, 26, 27) # define an index
ASD_National[ix] <- lapply(ASD_National[ix], as.factor)
```

Create new ordered factor Year_Factor from Year

In [75]:
```
ASD_National$Year_Factor <- factor(ASD_National$Year, ordered = TRUE)
```

In [76]:
```
# Observe the difference of 'Levels' in below two factors
ASD_National$Year_Factor # Ordinal categorical variable
str(ASD_National$Year_Factor)

ASD_National$Source # Nominal categorical variable
str(ASD_National$Source)
```

2000 2002 2004 2006 2008 2010 2012 2014 2004 2008 2012 2016 2000 2001
2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
2016 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012

▶ **Levels**:

```
 Ord.factor w/ 17 levels "2000"<"2001"<..: 1 3 5 7 9 11 13 15 5 9 ...
```

addm addm addm addm addm addm addm addm nsch nsch nsch nsch sped sped
sped sped sped sped sped sped sped sped sped sped sped sped sped sped
sped medi medi medi medi medi medi medi medi medi medi medi medi medi

▶ **Levels**:

```
 Factor w/ 4 levels "addm","medi",..: 1 1 1 1 1 1 1 1 3 3 ...
```

Convert Prevalence_Rank2 & Prevalence_Rank4 to ordered factor

In [77]:
```
# Convert to factor
ASD_National$Prevalence_Risk2 = factor(ASD_National$Prevalence_Risk2, ordered=
                                       levels=c("Low", "High"))
# Convert to factor
ASD_National$Prevalence_Risk4 = factor(ASD_National$Prevalence_Risk4, ordered=
                                       levels=c("Low", "Medium", "High", "
```

In [78]:
```
# Optionally, below is manual conversion examples:
# ASD_National$Male.Prevalence = as.numeric(ASD_National$Male.Prevalence)
# ASD_National$Source = as.factor(ASD_National$Source)
# ASD_National$Prevalence_Risk2 = factor(ASD_National$Prevalence_Risk2, ordere
# ASD_National$Prevalence_Risk4 = factor(ASD_National$Prevalence_Risk4, ordere
```

Optionally, export the processed dataframe data to CSV file.

In [79]:
```
write.csv(ASD_National, file = "../dataset/ADV_ASD_National_R.csv", row.names
```

In [80]:
```
# Read back in above saved file:
# ASD_National <- read.csv("../dataset/ADV_ASD_National_R.csv")
# ASD_National$Year_Factor <- factor(ASD_National$Year_Factor, ordered = TRUE)
```

Data Summarization

Data Summarization - High Level Data Summary

In [81]:
```
summary(ASD_National)
```

```
 Source        Year        Prevalence        Upper.CI         Lower.CI
addm: 8   Min.   :2000   Min.   : 1.800   Min.   : 1.800   Min.   : 1.700
medi:13   1st Qu.:2004   1st Qu.: 3.950   1st Qu.: 3.950   1st Qu.: 3.875
nsch: 4   Median :2008   Median : 6.650   Median : 6.900   Median : 6.350
sped:17   Mean   :2007   Mean   : 7.952   Mean   : 8.207   Mean   : 7.712
          3rd Qu.:2011   3rd Qu.: 9.725   3rd Qu.:10.350   3rd Qu.: 9.625
          Max.   :2016   Max.   :29.200   Max.   :30.700   Max.   :27.700

                                                 Source_Full1
Autism & Developmental Disabilities Monitoring Network: 8
Medicaid                                                :13
National Survey of Children's Health                    : 4
Special Education Child Count                           :17



                                                    Source_Full2
addm-Autism & Developmental Disabilities Monitoring Network: 8
medi-Medicaid                                                :13
nsch-National Survey of Children's Health                    : 4
sped-Special Education Child Count                           :17



Male.Prevalence Male.Lower.CI   Male.Upper.CI   Female.Prevalence
Min.   :11.50   Min.   :12.20   Min.   :13.70   Min.   :2.700
1st Qu.:13.70   1st Qu.:14.85   1st Qu.:16.07   1st Qu.:3.050
Median :18.40   Median :20.20   Median :21.55   Median :4.000
Mean   :18.71   Mean   :19.22   Mean   :20.62   Mean   :4.271
3rd Qu.:23.55   3rd Qu.:22.93   3rd Qu.:24.32   3rd Qu.:5.250
Max.   :26.60   Max.   :25.80   Max.   :27.40   Max.   :6.600
NA's   :35      NA's   :36      NA's   :36      NA's   :35
Female.Lower.CI Female.Upper.CI Non.hispanic.white.Prevalence
Min.   :2.600   Min.   :3.300   Min.   : 7.70
1st Qu.:3.100   1st Qu.:3.700   1st Qu.: 9.80
Median :4.300   Median :4.950   Median :12.00
Mean   :4.217   Mean   :4.900   Mean   :12.51
3rd Qu.:4.975   3rd Qu.:5.675   3rd Qu.:15.55
Max.   :6.200   Max.   :7.000   Max.   :17.20
NA's   :36      NA's   :36      NA's   :35
Non.hispanic.white.Lower.CI Non.hispanic.white.Upper.CI
Min.   : 9.100              Min.   :10.40
1st Qu.: 9.925              1st Qu.:10.93
Median :13.100              Median :14.20
Mean   :12.733              Mean   :13.88
3rd Qu.:15.075              3rd Qu.:16.20
Max.   :16.500              Max.   :17.80
NA's   :36                  NA's   :36
Non.hispanic.black.Prevalence Non.hispanic.black.Lower.CI
Min.   : 6.50                 Min.   : 6.200
1st Qu.: 7.05                 1st Qu.: 7.325
Median :10.20                 Median :10.500
Mean   :10.31                 Mean   :10.200
3rd Qu.:12.70                 3rd Qu.:12.100
Max.   :16.00                 Max.   :15.100
NA's   :35                    NA's   :36
Non.hispanic.black.Upper.CI Hispanic.Prevalence Hispanic.Lower.CI
Min.   : 7.600              Min.   : 5.900      Min.   : 5.000
1st Qu.: 8.575              1st Qu.: 6.625      1st Qu.: 5.775
Median :12.000              Median : 9.000      Median : 8.300
Mean   :11.700              Mean   : 9.150      Mean   : 8.333
3rd Qu.:13.700              3rd Qu.:10.625      3rd Qu.: 9.850
Max.   :16.900              Max.   :14.000      Max.   :13.100
```

```
NA's   :36                  NA's   :36          NA's   :36
Hispanic.Upper.CI Asian.or.Pacific.Islander.Prevalence
Min.   : 6.600    Min.   : 9.70
1st Qu.: 7.775    1st Qu.:10.97
Median : 9.750    Median :11.85
Mean   :10.017    Mean   :11.72
3rd Qu.:11.425    3rd Qu.:12.60
Max.   :14.900    Max.   :13.50
NA's   :36        NA's   :38
Asian.or.Pacific.Islander.Lower.CI Asian.or.Pacific.Islander.Upper.CI
Min.   : 8.10                      Min.   :11.60
1st Qu.: 9.45                      1st Qu.:12.72
Median :10.30                      Median :13.65
Mean   :10.12                      Mean   :13.57
3rd Qu.:10.97                      3rd Qu.:14.50
Max.   :11.80                      Max.   :15.40
NA's   :38                         NA's   :38
Source_UC                                                          Source_Full3
ADDM: 8   ADDM Autism & Developmental Disabilities Monitoring Network: 8
MEDI:13   MEDI Medicaid                                                    :13
NSCH: 4   NSCH National Survey of Children's Health                       : 4
SPED:17   SPED Special Education Child Count                             :17


Prevalence_Risk2  Prevalence_Risk4  Year_Factor
Low :14           Low      :14      2004   : 4
High:28           Medium   :18      2008   : 4
                  High     : 8      2012   : 4
                  Very High: 2      2000   : 3
                                    2002   : 3
                                    2006   : 3
                                    (Other):21
```

Data Summarization - Summary of numeric variables

In [82]:
```
# Filter only numeric variables/columns
select_if(ASD_National, is.numeric) # library(dplyr)
```

Year	Prevalence	Upper.CI	Lower.CI	Male.Prevalence	Male.Lower.CI	Male.Upper.CI	Female.Prevalence
2000	6.7	7.0	6.3	NA	NA	NA	NA
2002	6.6	6.8	6.3	11.5	NA	NA	2.7
2004	8.0	8.4	7.6	12.9	12.2	13.7	2.9
2006	9.0	9.3	8.6	14.5	13.9	15.1	3.2
2008	11.3	11.7	11.0	18.4	17.7	19.0	4.0
2010	14.7	15.1	14.3	23.7	23.0	24.4	5.3
2012	14.8	15.2	14.4	23.4	22.7	24.1	5.2
2014	16.8	17.3	16.4	26.6	25.8	27.4	6.6
2004	9.5	12.0	7.4	NA	NA	NA	NA
2008	16.2	18.1	14.5	NA	NA	NA	NA
2012	21.2	22.3	20.1	NA	NA	NA	NA
2016	29.2	30.7	27.7	NA	NA	NA	NA
2000	1.8	1.8	1.7	NA	NA	NA	NA
2001	2.1	2.1	2.1	NA	NA	NA	NA
2002	2.6	2.6	2.6	NA	NA	NA	NA
2003	3.0	3.0	3.0	NA	NA	NA	NA
2004	3.6	3.6	3.5	NA	NA	NA	NA
2005	4.1	4.1	4.1	NA	NA	NA	NA
2006	4.8	4.8	4.7	NA	NA	NA	NA
2007	5.4	5.5	5.4	NA	NA	NA	NA
2008	6.2	6.2	6.2	NA	NA	NA	NA
2009	7.0	7.0	7.0	NA	NA	NA	NA
2010	7.7	7.7	7.7	NA	NA	NA	NA
2011	8.4	8.5	8.4	NA	NA	NA	NA
2012	9.1	9.2	9.1	NA	NA	NA	NA
2013	9.8	9.9	9.8	NA	NA	NA	NA
2014	10.5	10.5	10.5	NA	NA	NA	NA
2015	11.2	11.2	11.2	NA	NA	NA	NA
2016	11.9	11.9	11.9	NA	NA	NA	NA
2000	2.3	2.4	2.3	NA	NA	NA	NA
2001	2.6	2.6	2.6	NA	NA	NA	NA
2002	2.8	2.8	2.7	NA	NA	NA	NA
2003	3.0	3.0	3.0	NA	NA	NA	NA
2004	3.5	3.6	3.5	NA	NA	NA	NA
2005	3.9	3.9	3.8	NA	NA	NA	NA
2006	4.4	4.5	4.4	NA	NA	NA	NA
2007	5.1	5.1	5.0	NA	NA	NA	NA
2008	5.6	5.6	5.5	NA	NA	NA	NA
2009	5.9	5.9	5.9	NA	NA	NA	NA

Year	Prevalence	Upper.CI	Lower.CI	Male.Prevalence	Male.Lower.CI	Male.Upper.CI	Female.Prevalence
2010	6.4	6.4	6.4	NA	NA	NA	NA
2011	7.1	7.1	7.1	NA	NA	NA	NA
2012	8.2	8.3	8.2	NA	NA	NA	NA

In [83]:
```
# Data summarization
summary(select_if(ASD_National, is.numeric))
```

```
      Year        Prevalence        Upper.CI         Lower.CI
 Min.   :2000   Min.   : 1.800   Min.   : 1.800   Min.   : 1.700
 1st Qu.:2004   1st Qu.: 3.950   1st Qu.: 3.950   1st Qu.: 3.875
 Median :2008   Median : 6.650   Median : 6.900   Median : 6.350
 Mean   :2007   Mean   : 7.952   Mean   : 8.207   Mean   : 7.712
 3rd Qu.:2011   3rd Qu.: 9.725   3rd Qu.:10.350   3rd Qu.: 9.625
 Max.   :2016   Max.   :29.200   Max.   :30.700   Max.   :27.700

 Male.Prevalence Male.Lower.CI   Male.Upper.CI   Female.Prevalence
 Min.   :11.50   Min.   :12.20   Min.   :13.70   Min.   :2.700
 1st Qu.:13.70   1st Qu.:14.85   1st Qu.:16.07   1st Qu.:3.050
 Median :18.40   Median :20.20   Median :21.55   Median :4.000
 Mean   :18.71   Mean   :19.22   Mean   :20.62   Mean   :4.271
 3rd Qu.:23.55   3rd Qu.:22.93   3rd Qu.:24.32   3rd Qu.:5.250
 Max.   :26.60   Max.   :25.80   Max.   :27.40   Max.   :6.600
 NA's   :35      NA's   :36      NA's   :36      NA's   :35
 Female.Lower.CI Female.Upper.CI Non.hispanic.white.Prevalence
 Min.   :2.600   Min.   :3.300   Min.   : 7.70
 1st Qu.:3.100   1st Qu.:3.700   1st Qu.: 9.80
 Median :4.300   Median :4.950   Median :12.00
 Mean   :4.217   Mean   :4.900   Mean   :12.51
 3rd Qu.:4.975   3rd Qu.:5.675   3rd Qu.:15.55
 Max.   :6.200   Max.   :7.000   Max.   :17.20
 NA's   :36      NA's   :36      NA's   :35
 Non.hispanic.white.Lower.CI Non.hispanic.white.Upper.CI
 Min.   : 9.100              Min.   :10.40
 1st Qu.: 9.925              1st Qu.:10.93
 Median :13.100              Median :14.20
 Mean   :12.733              Mean   :13.88
 3rd Qu.:15.075              3rd Qu.:16.20
 Max.   :16.500              Max.   :17.80
 NA's   :36                  NA's   :36
 Non.hispanic.black.Prevalence Non.hispanic.black.Lower.CI
 Min.   : 6.50                 Min.   : 6.200
 1st Qu.: 7.05                 1st Qu.: 7.325
 Median :10.20                 Median :10.500
 Mean   :10.31                 Mean   :10.200
 3rd Qu.:12.70                 3rd Qu.:12.100
 Max.   :16.00                 Max.   :15.100
 NA's   :35                    NA's   :36
 Non.hispanic.black.Upper.CI Hispanic.Prevalence Hispanic.Lower.CI
 Min.   : 7.600              Min.   : 5.900      Min.   : 5.000
 1st Qu.: 8.575              1st Qu.: 6.625      1st Qu.: 5.775
 Median :12.000              Median : 9.000      Median : 8.300
 Mean   :11.700              Mean   : 9.150      Mean   : 8.333
 3rd Qu.:13.700              3rd Qu.:10.625      3rd Qu.: 9.850
 Max.   :16.900              Max.   :14.000      Max.   :13.100
 NA's   :36                  NA's   :36          NA's   :36
 Hispanic.Upper.CI Asian.or.Pacific.Islander.Prevalence
 Min.   : 6.600    Min.   : 9.70
 1st Qu.: 7.775    1st Qu.:10.97
 Median : 9.750    Median :11.85
 Mean   :10.017    Mean   :11.72
 3rd Qu.:11.425    3rd Qu.:12.60
 Max.   :14.900    Max.   :13.50
 NA's   :36        NA's   :38
 Asian.or.Pacific.Islander.Lower.CI Asian.or.Pacific.Islander.Upper.CI
 Min.   : 8.10                      Min.   :11.60
 1st Qu.: 9.45                      1st Qu.:12.72
 Median :10.30                      Median :13.65
 Mean   :10.12                      Mean   :13.57
 3rd Qu.:10.97                      3rd Qu.:14.50
```

```
Max.   :11.80                    Max.   :15.40
NA's   :38                       NA's   :38
```

[Tips] We notice missing data in a few Prevalence variables.

In [84]:
```
# Calculate average Prevalence, no error
mean(ASD_National$Prevalence)
mean(ASD_National$Prevalence[ASD_National$Source == 'addm'])
mean(ASD_National$Prevalence[ASD_National$Source == 'medi'])
mean(ASD_National$Prevalence[ASD_National$Source == 'nsch'])
mean(ASD_National$Prevalence[ASD_National$Source == 'sped'])
```

7.95238095238095

10.9875

4.67692307692308

19.025

6.42352941176471

In [85]:
```
# Calculate average Male.Prevalence, there is error!
mean(ASD_National$Male.Prevalence)
```

<NA>

In [86]:
```
# Because of NA, mean() cannot process, thus we use na.rm to ignore NAs
mean(ASD_National$Male.Prevalence, na.rm = TRUE)
```

18.7142857142857

In [87]:
```
mean(ASD_National$Female.Prevalence, na.rm = TRUE)
```

4.27142857142857

In [88]:
```
# Count occurrences of uniques values in a variable/column: number of rows (of
table(ASD_National$Year) # ?table
```

```
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 20
15
   3    2    3    2    4    2    3    2    4    2    3    2    4    1    2
1
2016
   2
```

Data Summarization - Summary of categorical variables

In [89]:
```
# List of categorical variables
names(select_if(ASD_National, is.factor)) # All categorical variables are fact
names(select_if(ASD_National, is.character)) # No categorical variable is char
```

'Source' 'Source_Full1' 'Source_Full2' 'Source_UC' 'Source_Full3' 'Prevalence_Risk2' 'Prevalence_Risk4' 'Year_Factor'

In [90]:
```
# Look at summary
summary(select_if(ASD_National, is.factor))
```

```
 Source                                                    Source_Full1
addm: 8   Autism & Developmental Disabilities Monitoring Network: 8
medi:13   Medicaid                                                :13
nsch: 4   National Survey of Children's Health                    : 4
sped:17   Special Education Child Count                           :17



                                                        Source_Full2 Source_UC
addm-Autism & Developmental Disabilities Monitoring Network: 8    ADDM: 8
medi-Medicaid                                                :13    MEDI:13
nsch-National Survey of Children's Health                    : 4    NSCH: 4
sped-Special Education Child Count                           :17    SPED:17



                                                        Source_Full3
ADDM Autism & Developmental Disabilities Monitoring Network: 8
MEDI Medicaid                                                :13
NSCH National Survey of Children's Health                    : 4
SPED Special Education Child Count                           :17



Prevalence_Risk2  Prevalence_Risk4  Year_Factor
Low :14           Low      :14      2004   : 4
High:28           Medium   :18      2008   : 4
                  High     : 8      2012   : 4
                  Very High: 2      2000   : 3
                                    2002   : 3
                                    2006   : 3
                                    (Other):21
```

In [91]:
```
summary(select_if(ASD_National, is.character))
```

```
< table of extent 0 x 0 >
```

In [92]:
```
# Count occurrences of uniques values in a variable/column
table(ASD_National$Source)
```

```
addm medi nsch sped
   8   13    4   17
```

In [93]:
```
table(ASD_National$Source_Full3)
```

```
ADDM Autism & Developmental Disabilities Monitoring Network
                                                           8
                                                MEDI Medicaid
                                                          13
                                NSCH National Survey of Children's Health
                                                           4
                                       SPED Special Education Child Count
                                                          17
```

In [94]:
```
table(ASD_National$Year_Factor)
```

```
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 20
15
   3    2    3    2    4    2    3    2    4    2    3    2    4    1    2
1
2016
   2
```

In [95]:
```
table(ASD_National$Prevalence) # numeric is also possible
```

```
 1.8  2.1  2.3  2.6  2.8    3  3.5  3.6  3.9  4.1  4.4  4.8  5.1  5.4  5.6
5.9
   1    1    1    2    1    2    1    1    1    1    1    1    1    1    1
1
 6.2  6.4  6.6  6.7    7  7.1  7.7    8  8.2  8.4    9  9.1  9.5  9.8 10.5 1
1.2
   1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
1
11.3 11.9 14.7 14.8 16.2 16.8 21.2 29.2
   1    1    1    1    1    1    1    1
```

In [96]:
```
# Display unique values (levels) of a factor categorical
lapply(select_if(ASD_National, is.factor), levels)
```

$Source
'addm' 'medi' 'nsch' 'sped'

$Source_Full1
'Autism & Developmental Disabilities Monitoring Network' 'Medicaid'
'National Survey of Children\'s Health' 'Special Education Child Count'

$Source_Full2
'addm-Autism & Developmental Disabilities Monitoring Network' 'medi-Medicaid'
'nsch-National Survey of Children\'s Health' 'sped-Special Education Child Count'

$Source_UC
'ADDM' 'MEDI' 'NSCH' 'SPED'

$Source_Full3
'ADDM Autism & Developmental Disabilities Monitoring Network' 'MEDI Medicaid'
'NSCH National Survey of Children\'s Health' 'SPED Special Education Child Count'

$Prevalence_Risk2
'Low' 'High'

$Prevalence_Risk4
'Low' 'Medium' 'High' 'Very High'

$Year_Factor
'2000' '2001' '2002' '2003' '2004' '2005' '2006' '2007' '2008' '2009' '2010' '2011' '2012'
'2013' '2014' '2015' '2016'

In [97]:
```
# or using variable names
lapply(ASD_National[c('Source_UC', 'Year_Factor')], levels)
```

$Source_UC

'ADDM' 'MEDI' 'NSCH' 'SPED'

$Year_Factor

'2000' '2001' '2002' '2003' '2004' '2005' '2006' '2007' '2008' '2009' '2010' '2011' '2012' '2013' '2014' '2015' '2016'

In [98]:
```
# Pivot of counting occurrences
table(ASD_National$Source_Full3, ASD_National$Year) # table(ASD_National$Year,
```

	2000	2001	2002
ADDM Autism & Developmental Disabilities Monitoring Network	1	0	1
MEDI Medicaid	1	1	1
NSCH National Survey of Children's Health	0	0	0
SPED Special Education Child Count	1	1	1

	2003	2004	2005
ADDM Autism & Developmental Disabilities Monitoring Network	0	1	0
MEDI Medicaid	1	1	1
NSCH National Survey of Children's Health	0	1	0
SPED Special Education Child Count	1	1	1

	2006	2007	2008
ADDM Autism & Developmental Disabilities Monitoring Network	1	0	1
MEDI Medicaid	1	1	1
NSCH National Survey of Children's Health	0	0	1
SPED Special Education Child Count	1	1	1

	2009	2010	2011
ADDM Autism & Developmental Disabilities Monitoring Network	0	1	0
MEDI Medicaid	1	1	1
NSCH National Survey of Children's Health	0	0	0
SPED Special Education Child Count	1	1	1

	2012	2013	2014
ADDM Autism & Developmental Disabilities Monitoring Network	1	0	1
MEDI Medicaid	1	0	0
NSCH National Survey of Children's Health	1	0	0
SPED Special Education Child Count	1	1	1

	2015	2016
ADDM Autism & Developmental Disabilities Monitoring Network	0	0
MEDI Medicaid	0	0
NSCH National Survey of Children's Health	0	1
SPED Special Education Child Count	1	1

In [99]:
```
# Pivot of counting occurrences
table(ASD_National$Prevalence_Risk2, ASD_National$Source)

# Pivot of counting occurrences
table(ASD_National$Prevalence_Risk4, ASD_National$Source)
```

```
      addm medi nsch sped
  Low    0    7    0    7
  High   8    6    4   10
```

```
           addm medi nsch sped
  Low         0    7    0    7
  Medium      4    6    1    7
  High        4    0    1    3
  Very High   0    0    2    0
```

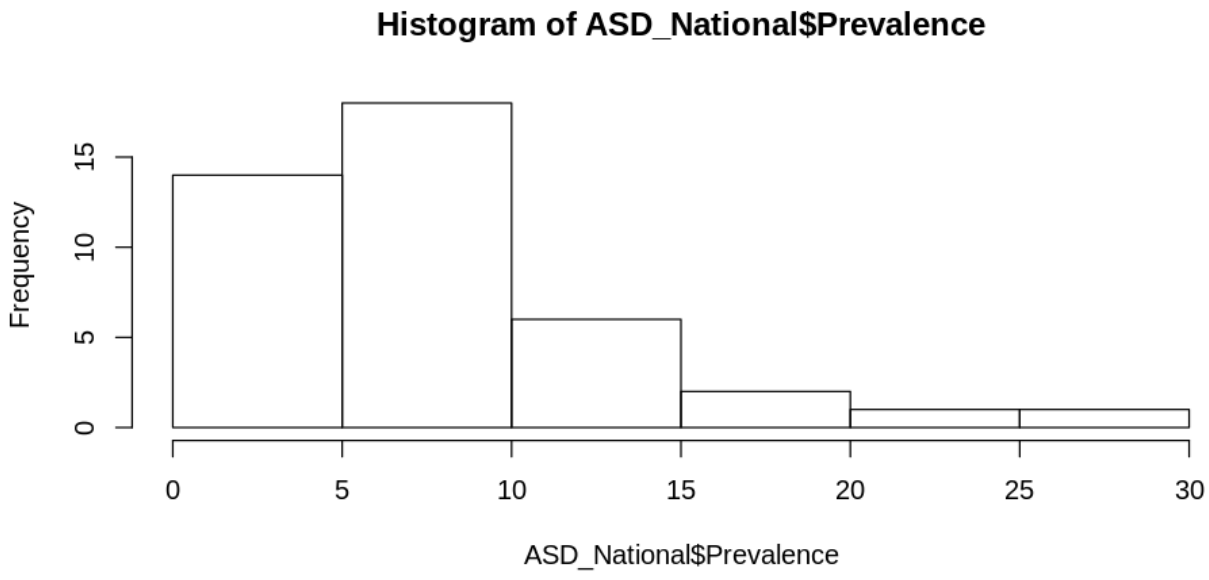
Data Visualisation (Base Graphic)

In [100]:
```
# library(repr)
# Adjust in-line plot size to M x N
options(repr.plot.width=8, repr.plot.height=4)
```

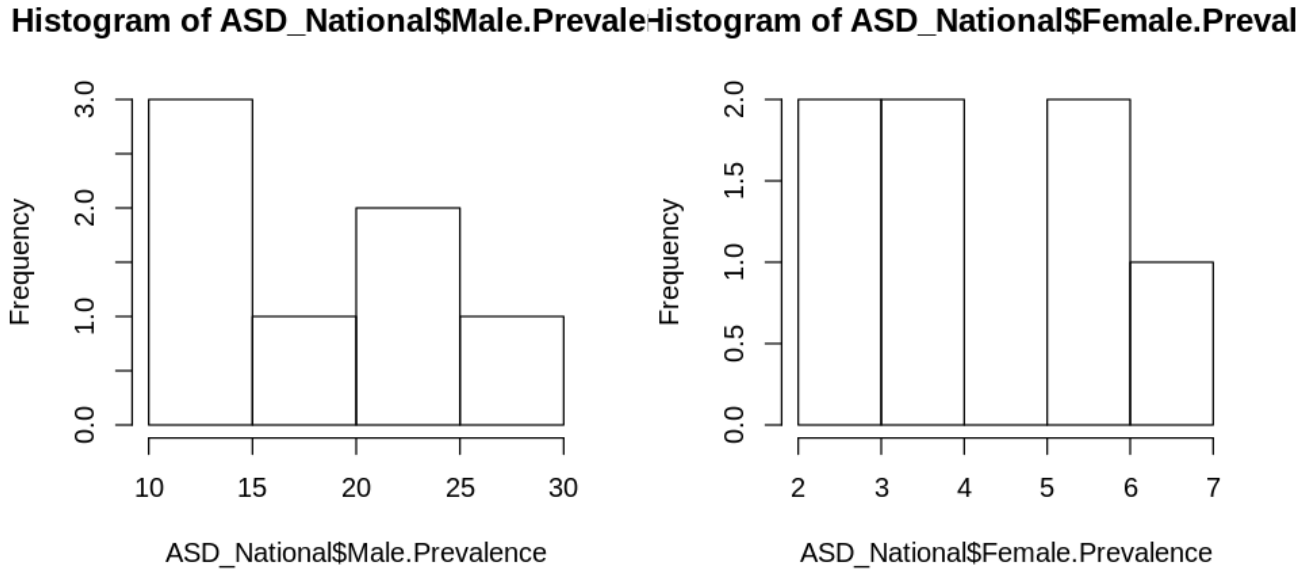
Data Visualisation (Base Graphic) - Histogram (distribution of binned continuous variable)

https://www.statmethods.net/graphs/density.html (https://www.statmethods.net/graphs/density.html)

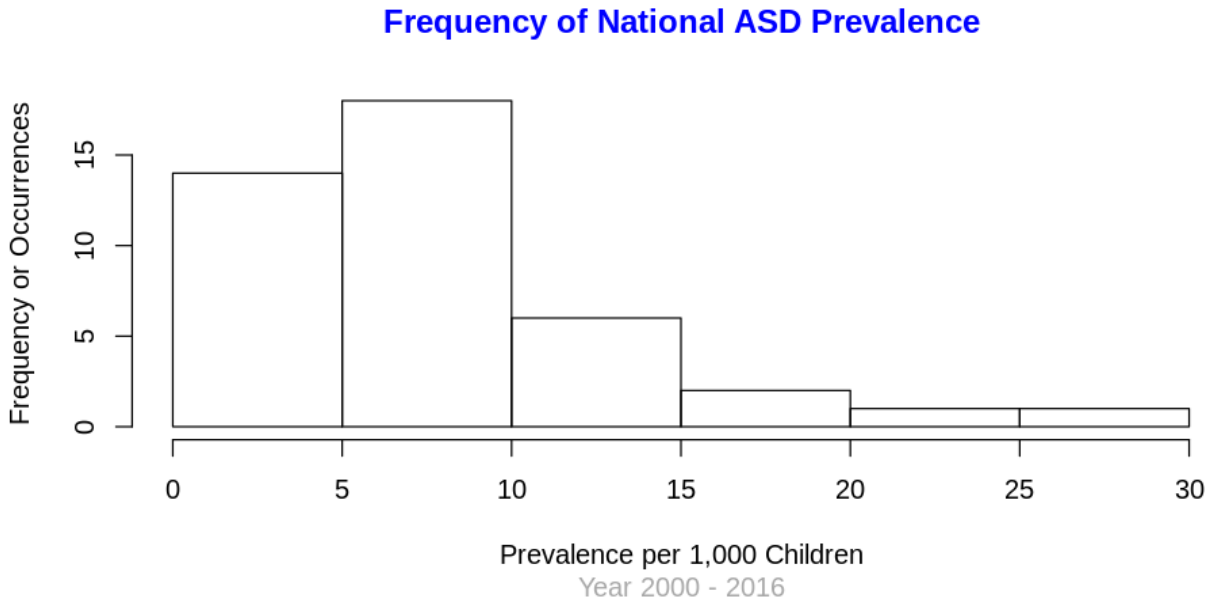
In [101]:
```
hist(ASD_National$Prevalence)
```



In [102]:
```
par(mfrow=c(1, 2)) # multiple plots on one page: row split to: 1,column split
hist(ASD_National$Male.Prevalence)
hist(ASD_National$Female.Prevalence)
par(mfrow=c(1, 1)) # Reset to one plot on one page
```



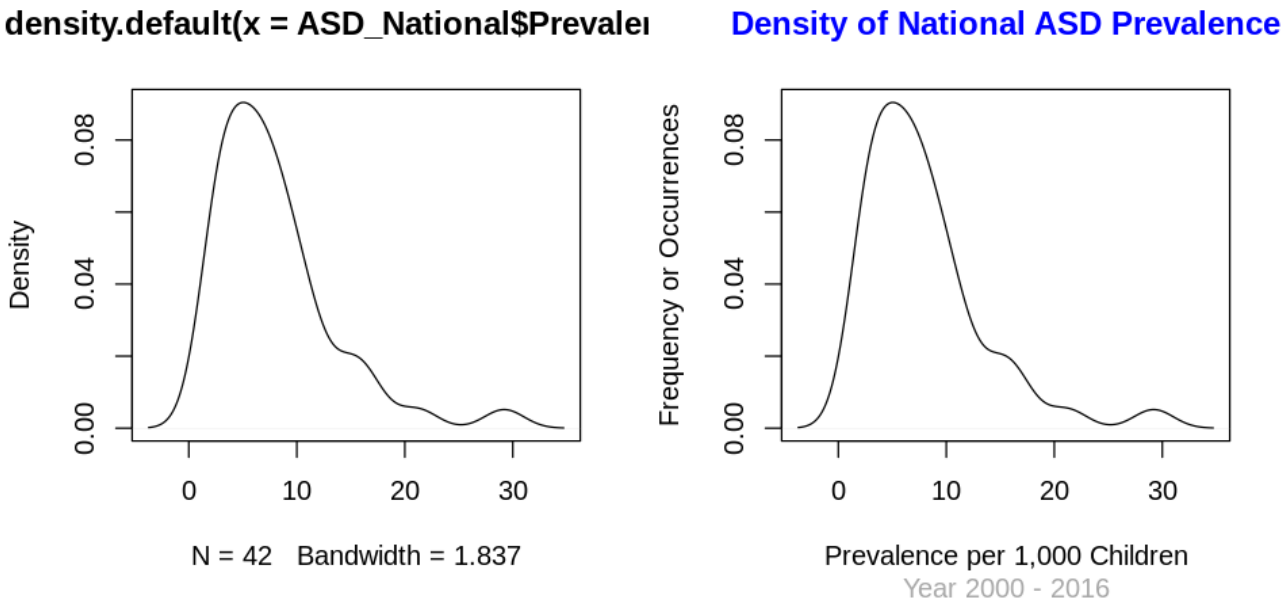
In [103]:
```
# Histogram with annotations
hist(ASD_National$Prevalence,
     main = "Frequency of National ASD Prevalence", # Chart title
     xlab = "Prevalence per 1,000 Children", # x axis label
     ylab = "Frequency or Occurrences",# y axis label
     sub  = "Year 2000 - 2016", # Chart subtitle at bottom
     col.main="blue", col.lab="black", col.sub="darkgrey") # Colours
```



Density plot (distribution for continuous variable normalized to 100% area under curve)

https://www.statmethods.net/graphs/density.html (https://www.statmethods.net/graphs/density.html)

In [104]:
```
par(mfrow=c(1, 2)) # multiple plots on one page: row split to: 1,column split

plot(density(ASD_National$Prevalence))

# Density plot with annotations
plot(density(ASD_National$Prevalence),
     main = "Density of National ASD Prevalence",
     xlab = "Prevalence per 1,000 Children",
     ylab = "Frequency or Occurrences",
     sub  = "Year 2000 - 2016",
     col.main="blue", col.lab="black", col.sub="darkgrey")

par(mfrow=c(1, 1))
```



Boxplot plot (median, 25% quantile,75% quantile)

https://www.statmethods.net/graphs/boxplot.html (https://www.statmethods.net/graphs/boxplot.html)

https://stats.stackexchange.com/questions/156778/percentile-vs-quantile-vs-quartile (https://stats.stackexchange.com/questions/156778/percentile-vs-quantile-vs-quartile)

0 quartile = 0 quantile = 0 percentile

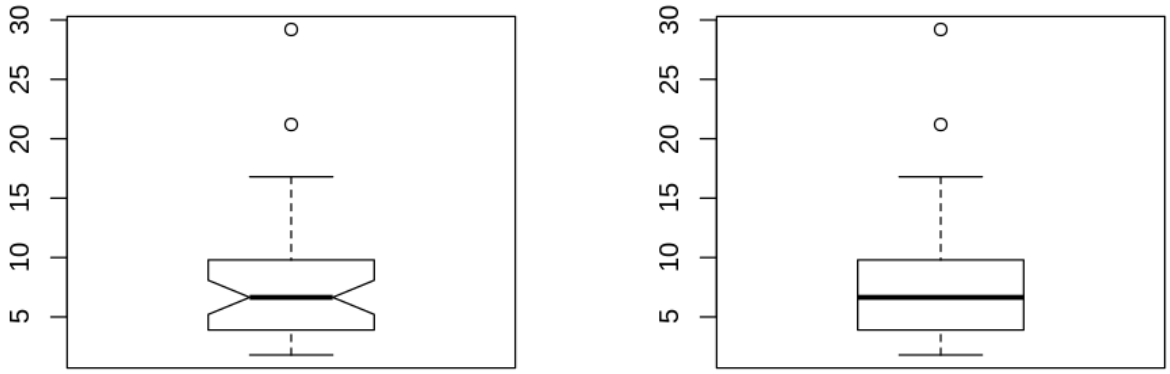
1 quartile = 0.25 quantile = 25 percentile

2 quartile = .5 quantile = 50 percentile (median)

3 quartile = .75 quantile = 75 percentile

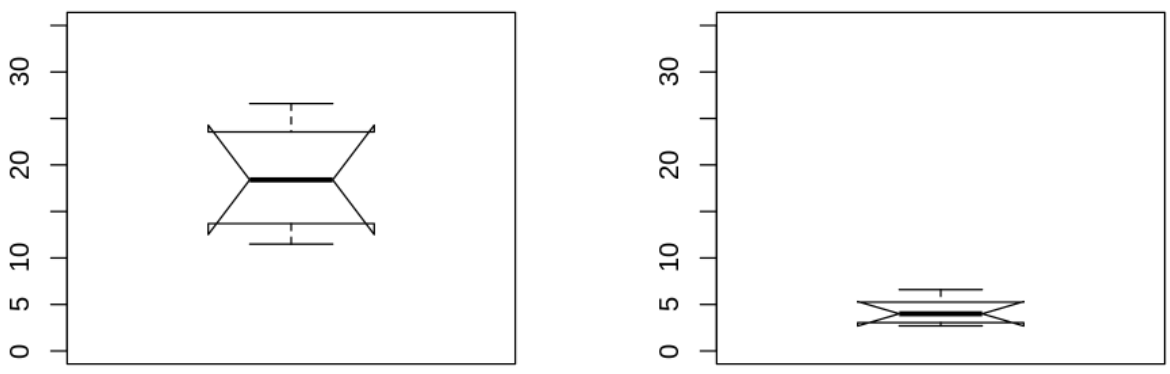
4 quartile = 1 quantile = 100 percentile

In [105]:
```
par(mfrow=c(1, 2)) # multiple plots on one page: row split to: 1,column split

# All children prevalence with and without 95% confidence side by side:
boxplot(ASD_National$Prevalence, notch = TRUE) # 95% confidence interval - a n
boxplot(ASD_National$Prevalence) # All children

par(mfrow=c(1, 1))
```



In [106]:
```
par(mfrow=c(1, 2)) # multiple plots on one page: row split to: 1,column split

# Male prevalence and Female prevalence side by side:
boxplot(ASD_National$Male.Prevalence, ylim = c(0, 35), notch = TRUE) # Male ch
boxplot(ASD_National$Female.Prevalence, ylim = c(0, 35), notch = TRUE) # Femal

par(mfrow=c(1, 1))
```

```
Warning message in bxp(list(stats = structure(c(11.5, 13.7, 18.4, 23.55, 26.
6), .Dim = c(5L, :
“some notches went outside hinges ('box'): maybe set notch=FALSE”Warning mess
age in bxp(list(stats = structure(c(2.7, 3.05, 4, 5.25, 6.6), .Dim = c(5L, :
“some notches went outside hinges ('box'): maybe set notch=FALSE”
```



In [107]:
```
# Display value ranges
# numeric:
range(ASD_National$Prevalence)
```

1.8 29.2

In [108]:
```
range(ASD_National$Year)
```

2000 2016

In [109]:
```
# categorical:
min(ASD_National$Year_Factor)
```

2000

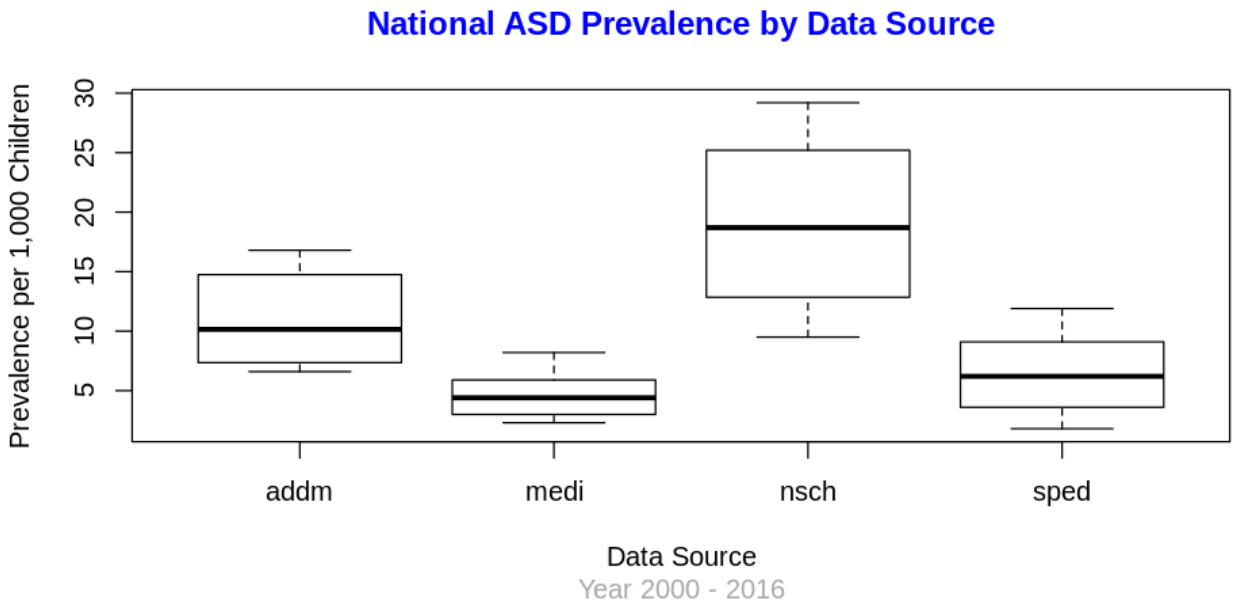
▶ **Levels**:

In [110]:
```
max(ASD_National$Year_Factor)
```

2016

▶ **Levels**:

In [111]:
```
# Create 'Prevalence' box plots break by 'Source'
boxplot(ASD_National$Prevalence ~ ASD_National$Source,
        main = "National ASD Prevalence by Data Source",
        xlab = "Data Source",
        ylab = "Prevalence per 1,000 Children",
        sub  = "Year 2000 - 2016",
        col.main="blue", col.lab="black", col.sub="darkgrey")
```



Quiz:

Set noth=TRUE to above boxplot. Are there overlapping among four data sources?

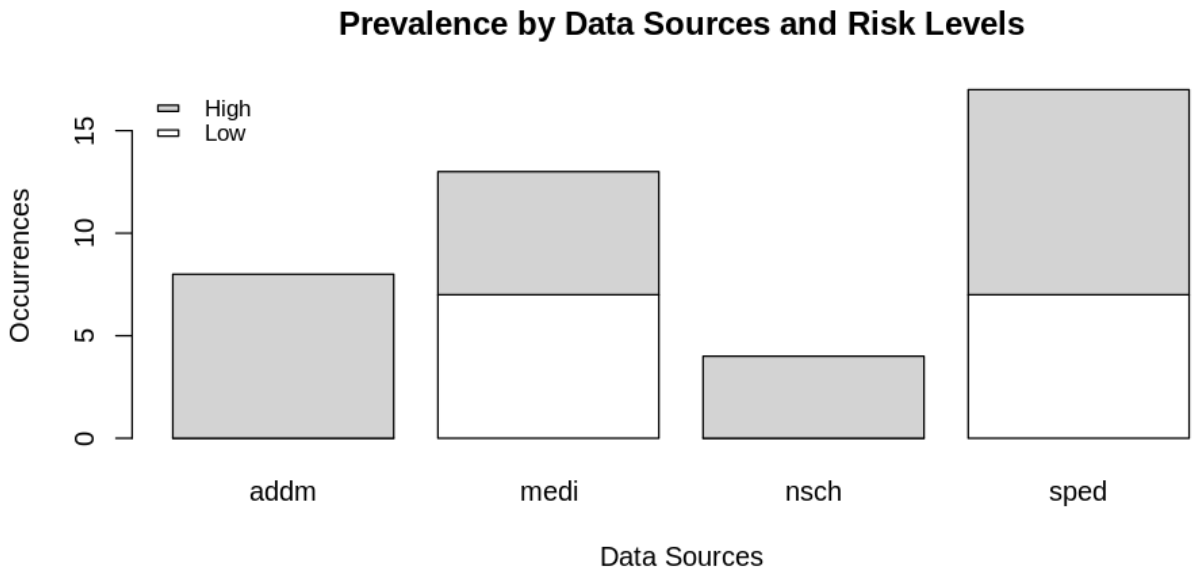
In [112]:
```
# Write your code below and press Shift+Enter to execute
```

Double-click **here** for the solution.

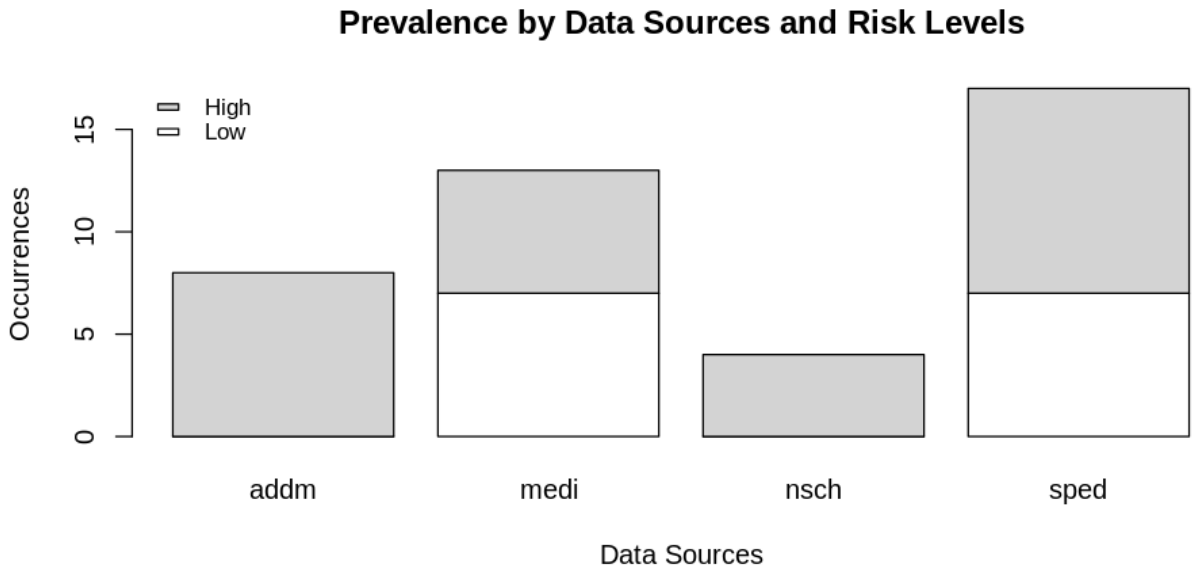
Data Visualisation (Base Graphic) - Bar plot

In [113]:
```
# Adjust in-line plot size to M x N
options(repr.plot.width=8, repr.plot.height=4)
```

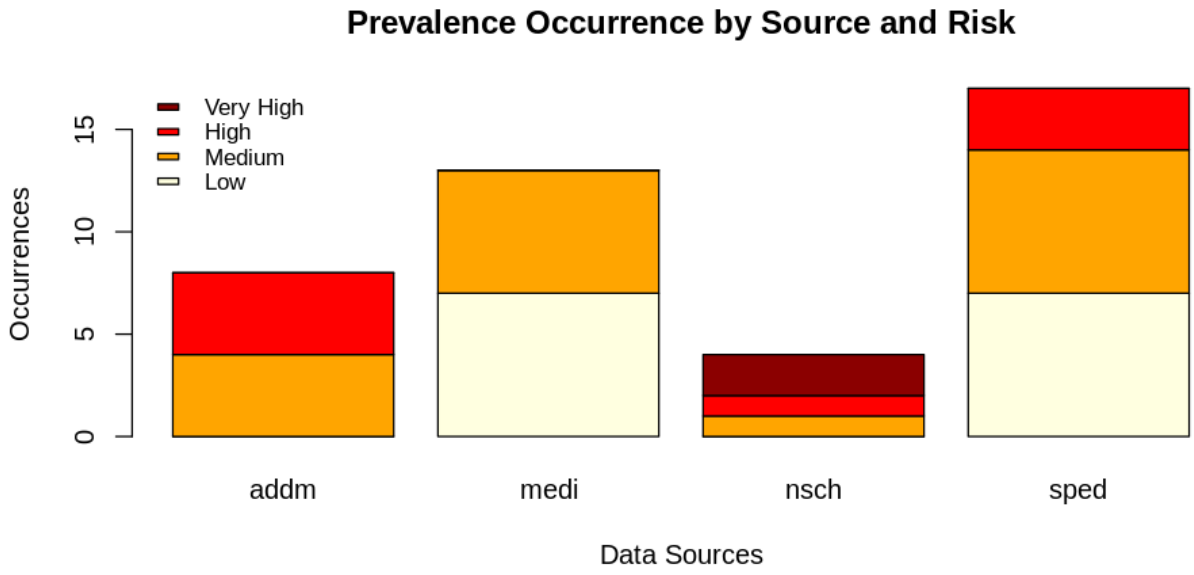
In [114]:
```
# ---------------------------------
# [National] Risk by Data Source
# ---------------------------------
# Create bar chart using R graphics
counts = table(ASD_National$Prevalence_Risk2, ASD_National$Source)
#counts = table(ASD_National$Source, ASD_National$Prevalence_Risk4)
barplot(counts,
        main="Prevalence by Data Sources and Risk Levels",
        xlab="Data Sources", col=c("white", "lightgrey"),
        ylab="Occurrences",
        legend = rownames(counts),
        args.legend = list(x="topleft", bty = "n", cex = 0.85, y.intersp=2))
```



In [115]:
```r
# ---------------------------------
# [National] Risk by Data Source
# ---------------------------------
# Create bar chart using R graphics
counts = table(ASD_National$Prevalence_Risk2, ASD_National$Source) # Count of
barplot(counts,
        main="Prevalence by Data Sources and Risk Levels",
        xlab="Data Sources",
        ylab="Occurrences",
        col=c("white", "lightgrey"),
        legend = rownames(counts),
        args.legend = list(x = "topleft", bty = "n", cex = 0.85, y.intersp = 2
```



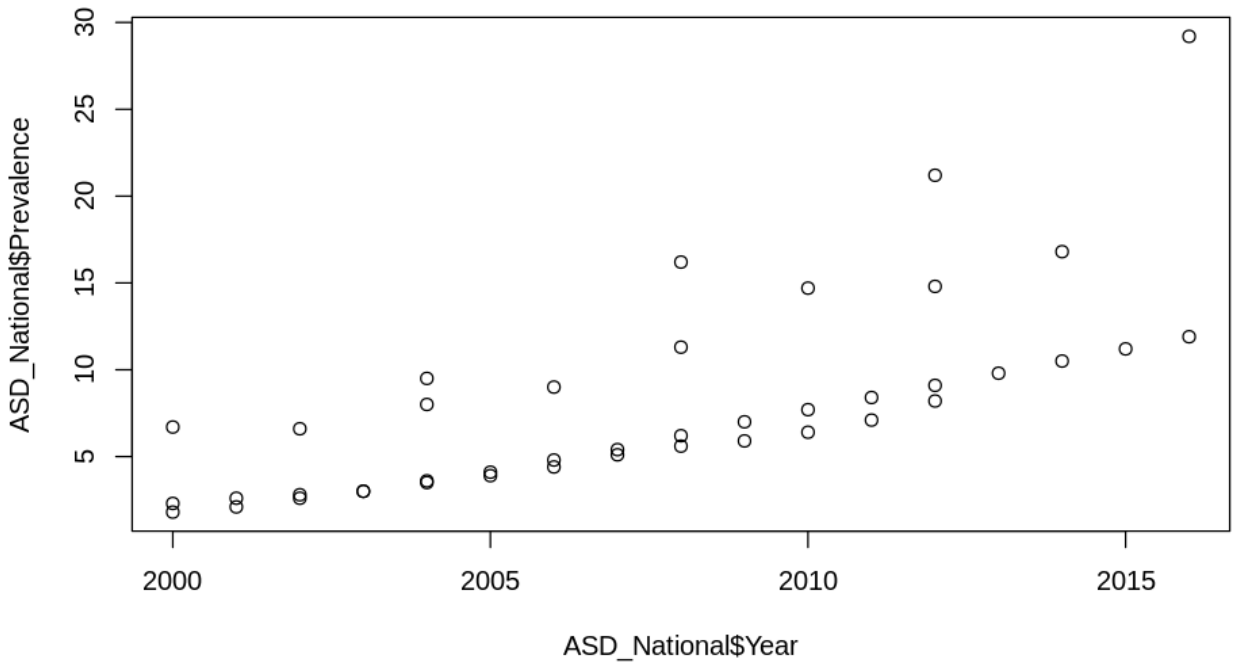
In [116]:
```
# ---------------------------------
# [National] Risk by Data Source
# ---------------------------------
# Create bar chart using R graphics
counts = table(ASD_National$Prevalence_Risk4, ASD_National$Source) # Count of
barplot(counts,
        main="Prevalence Occurrence by Source and Risk",
        xlab="Data Sources",
        ylab="Occurrences",
        col=c("lightyellow", "orange", "red","darkred"),
        legend = rownames(counts),
        args.legend = list(x = "topleft", bty = "n", cex = 0.85, y.intersp = 2
```



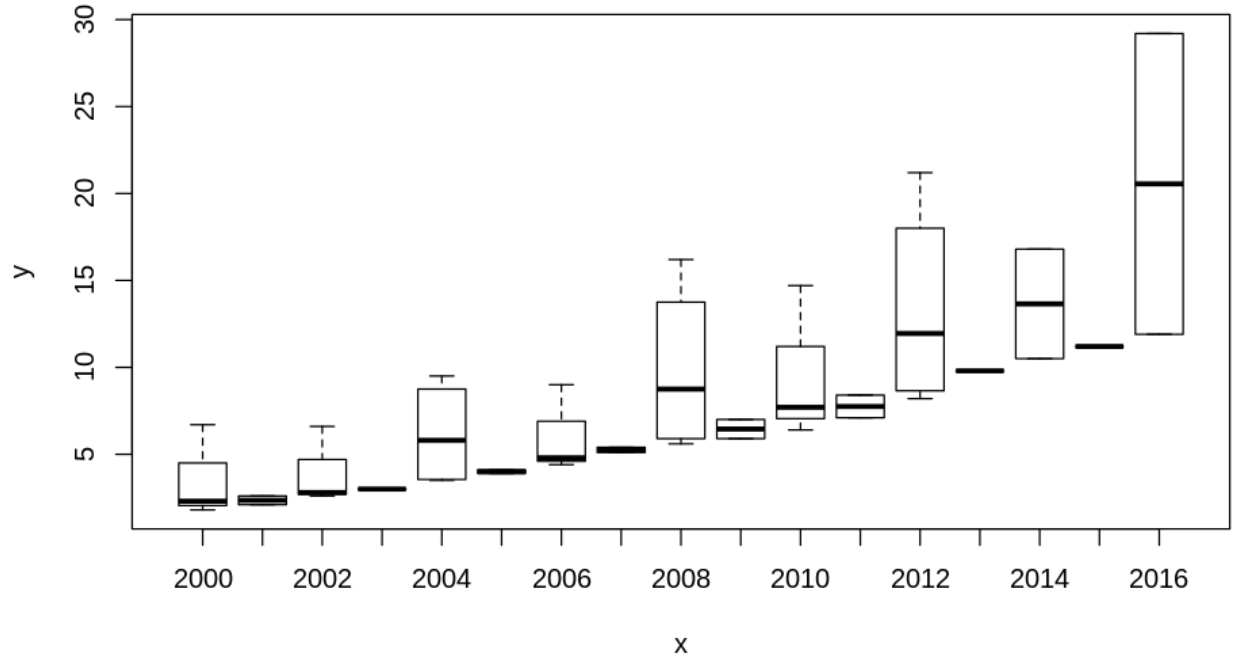
Data Visualisation (Base Graphic) - Line chart

In [117]:
```
# Adjust in-line plot size to M x N
options(repr.plot.width=8, repr.plot.height=5)
```

In [118]:
```
# ---------------------------------
# [National] < Prevalence has changed over Time >
# ---------------------------------
# Prevalence over Year
# Use Year        as x-axis: y value Prevalence is NOT aggregated for differen
plot(ASD_National$Year, ASD_National$Prevalence)
```



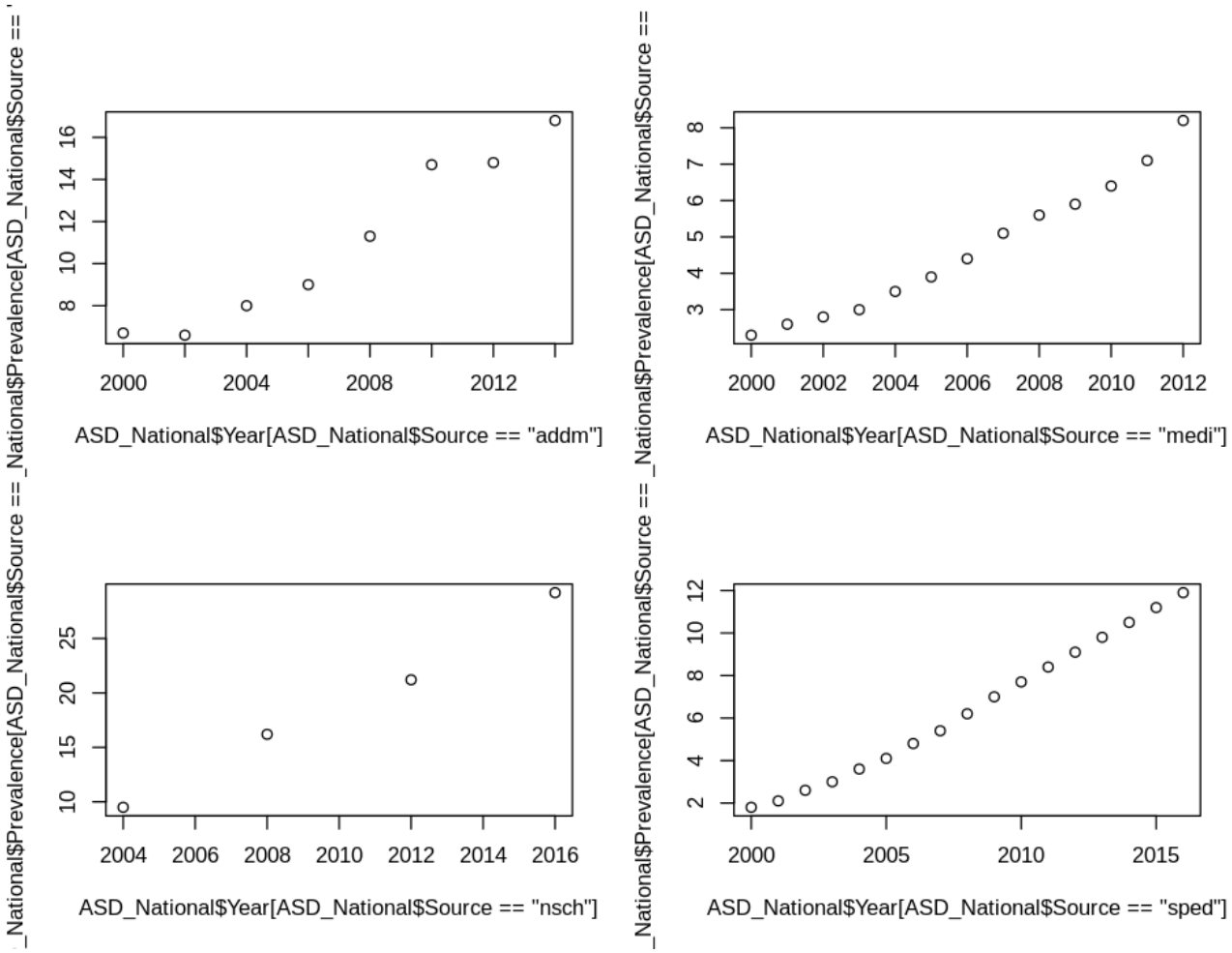
In [119]:
```
# Use Year_factor as x-axis: y value Prevalence is     aggregated for differen
plot(ASD_National$Year_Factor, ASD_National$Prevalence)
```



In [120]:
```
# table(ASD_National$Source_Full3)
```

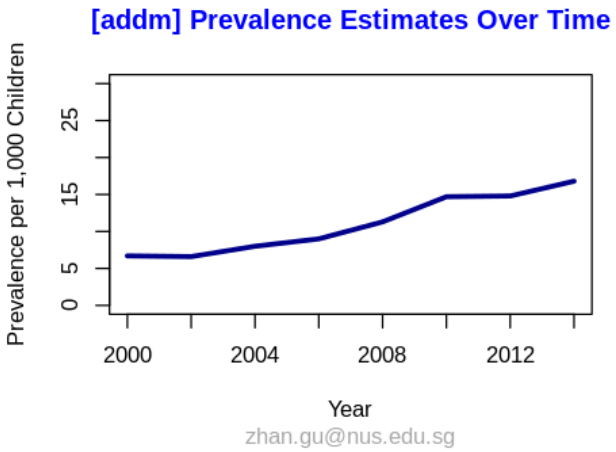
In [121]:

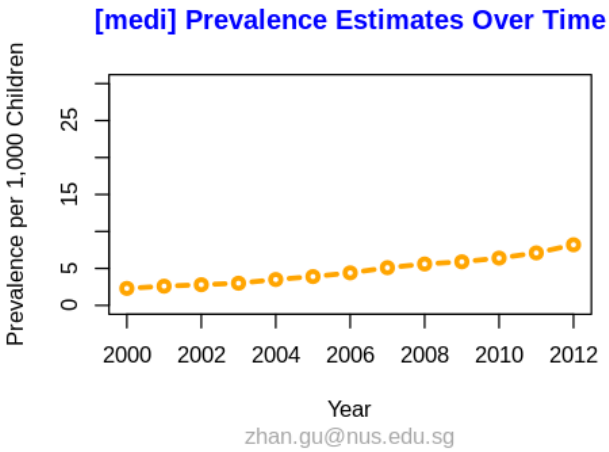
```
# Adjust in-line plot size to M x N
options(repr.plot.width=8, repr.plot.height=6)

par(mfrow=c(2, 2))

# Prevalence over Year, from data source:
# addm-Autism & Developmental Disabilities Monitoring Network
plot(ASD_National$Year[ASD_National$Source == 'addm'],
     ASD_National$Prevalence[ASD_National$Source == 'addm'])

# Prevalence over Year, from data source:
# medi-Medicaid
plot(ASD_National$Year[ASD_National$Source == 'medi'],
     ASD_National$Prevalence[ASD_National$Source == 'medi'])

# Prevalence over Year, from data source:
# nsch-National Survey of Children Health
plot(ASD_National$Year[ASD_National$Source == 'nsch'],
     ASD_National$Prevalence[ASD_National$Source == 'nsch'])

# Prevalence over Year, from data source:
# sped-Special Education Child Count
plot(ASD_National$Year[ASD_National$Source == 'sped'],
     ASD_National$Prevalence[ASD_National$Source == 'sped'])

par(mfrow=c(1, 1)) # Reset to one plot on one page
```

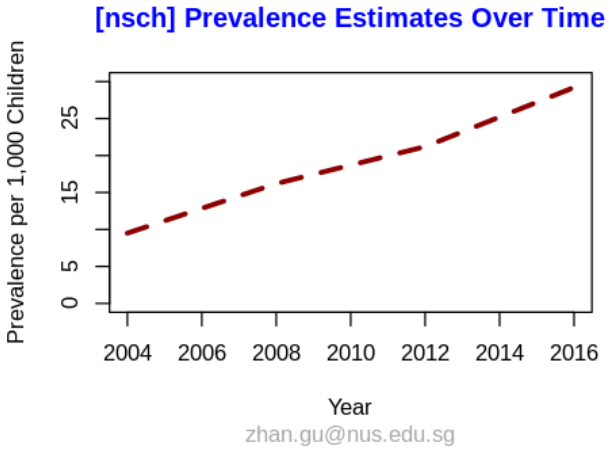


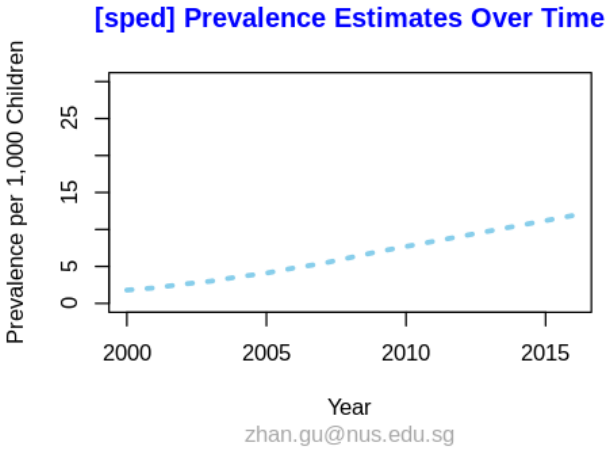
In [122]:

```
# ---------------------------------
# Add more annotations to above plots
# ---------------------------------
# Color list
# addm : darkblue
# medi : orange
# nsch : darkred
# sped : skyblue

par(mfrow=c(2, 2))

# Prevalence over Year, from data source:
# addm-Autism & Developmental Disabilities Monitoring Network
plot(ASD_National$Year[ASD_National$Source == 'addm'],
     ASD_National$Prevalence[ASD_National$Source == 'addm'],
     type="l", # dot/point type
     lty=1, # line type
     lwd=3, # line width
     col="darkblue", # line color
     xlab="Year",
     ylab="Prevalence per 1,000 Children",
     ylim = c(0, 30), # Set value range of y axis
     main="[addm] Prevalence Estimates Over Time",
     sub  = "zhan.gu@nus.edu.sg",
     col.main="blue", col.lab="black", col.sub="darkgrey")

# Prevalence over Year, from data source:
# medi-Medicaid
plot(ASD_National$Year[ASD_National$Source == 'medi'],
     ASD_National$Prevalence[ASD_National$Source == 'medi'],
     type="b", lty=1, lwd=3,  col="orange",
     xlab="Year",
     ylab="Prevalence per 1,000 Children",
     ylim = c(0, 30), # Set value range of y axis
     main="[medi] Prevalence Estimates Over Time",
     sub  = "zhan.gu@nus.edu.sg",
     col.main="blue", col.lab="black", col.sub="darkgrey")

# Prevalence over Year, from data source:
# nsch-National Survey of Children Health
plot(ASD_National$Year[ASD_National$Source == 'nsch'],
     ASD_National$Prevalence[ASD_National$Source == 'nsch'],
     type="l", lty=2, lwd=3,  col="darkred",
     xlab="Year",
     ylab="Prevalence per 1,000 Children",
     ylim = c(0, 30), # Set value range of y axis
     main="[nsch] Prevalence Estimates Over Time",
     sub  = "zhan.gu@nus.edu.sg",
     col.main="blue", col.lab="black", col.sub="darkgrey")

# Prevalence over Year, from data source:
# sped-Special Education Child Count
plot(ASD_National$Year[ASD_National$Source == 'sped'],
     ASD_National$Prevalence[ASD_National$Source == 'sped'],
     type="l", lty=3, lwd=3,  col="skyblue",
     xlab="Year",
     ylab="Prevalence per 1,000 Children",
     ylim = c(0, 30), # Set value range of y axis
     main="[sped] Prevalence Estimates Over Time",
     sub  = "zhan.gu@nus.edu.sg",
     col.main="blue", col.lab="black", col.sub="darkgrey")

par(mfrow=c(1, 1)) # Reset to one plot on one page
```







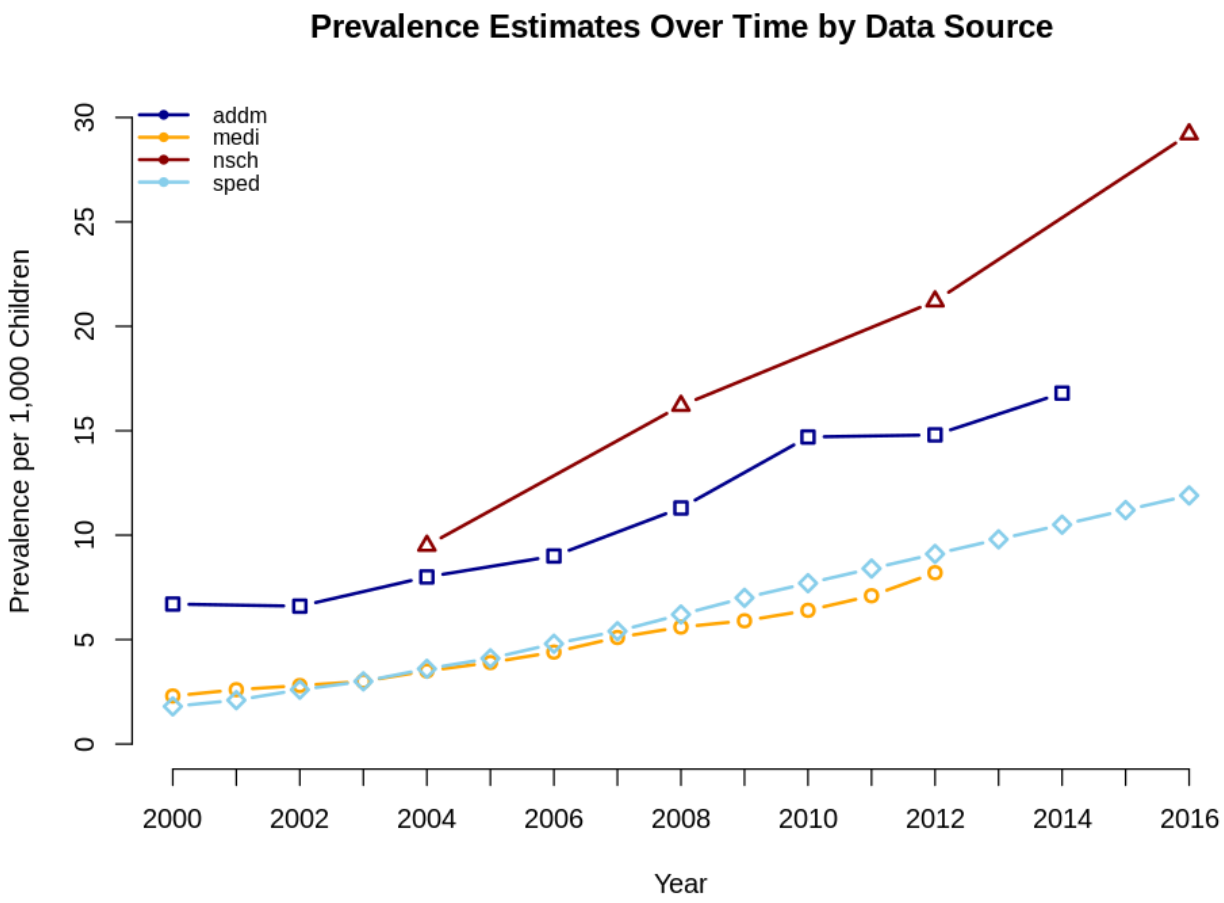


Data Visualisation (Base Graphic) - [R] REPORTED PREVALENCE HAS CHANGED OVER TIME by [Data Source]

Create multiple lines within a single chart

In [123]:

```r
# ---------------------------------
# [National] < Prevalence Varies over Time/Year by Data Source >
# ---------------------------------
# Create a first line
plot(ASD_National$Year[ASD_National$Source == 'addm'],
     ASD_National$Prevalence[ASD_National$Source == 'addm'],
     col = "darkblue", lty = 1, lwd = 2,
     type = "b", # use dot/point
     pch = 0, # dot/point type: http://www.endmemo.com/program/R/pchsymbols.ph
     xlab="Year",
     xlim=c(2000, 2016), # Set x axis value range
     ylab="Prevalence per 1,000 Children",
     ylim=c(0, 30), # Set y axis value range
     main="Prevalence Estimates Over Time by Data Source",
     col.main="black", col.lab="black", col.sub="grey",
     frame = FALSE, # Remove frame
     axes=FALSE # Remove x and y axis
)
axis(1, at=seq(2000, 2016, 1)) # Customize x axis
axis(2, at=seq(0, 30, 5)) # Customize y axis

# Add another line
lines(ASD_National$Year[ASD_National$Source == 'medi'],
      ASD_National$Prevalence[ASD_National$Source == 'medi'],
      pch = 1, col = "orange", type = "b", lty = 1, lwd = 2
)
# Add another line
lines(ASD_National$Year[ASD_National$Source == 'nsch'],
      ASD_National$Prevalence[ASD_National$Source == 'nsch'],
      pch = 2, col = "darkred", type = "b", lty = 1, lwd = 2
)
# Add another line
lines(ASD_National$Year[ASD_National$Source == 'sped'],
      ASD_National$Prevalence[ASD_National$Source == 'sped'],
      pch = 5, col = "skyblue", type = "b", lty = 1, lwd = 2
)
# Add a legend to the plot
legend("topleft", legend=levels(ASD_National$Source),
       col=c("darkblue", "orange", "darkred", "skyblue"),
       pch = 20, # dot in a line
       lty = 1, # line type
       lwd = 2, # line width
       cex=0.8, # size of text
       bty = 'n' # Without frame
)
```



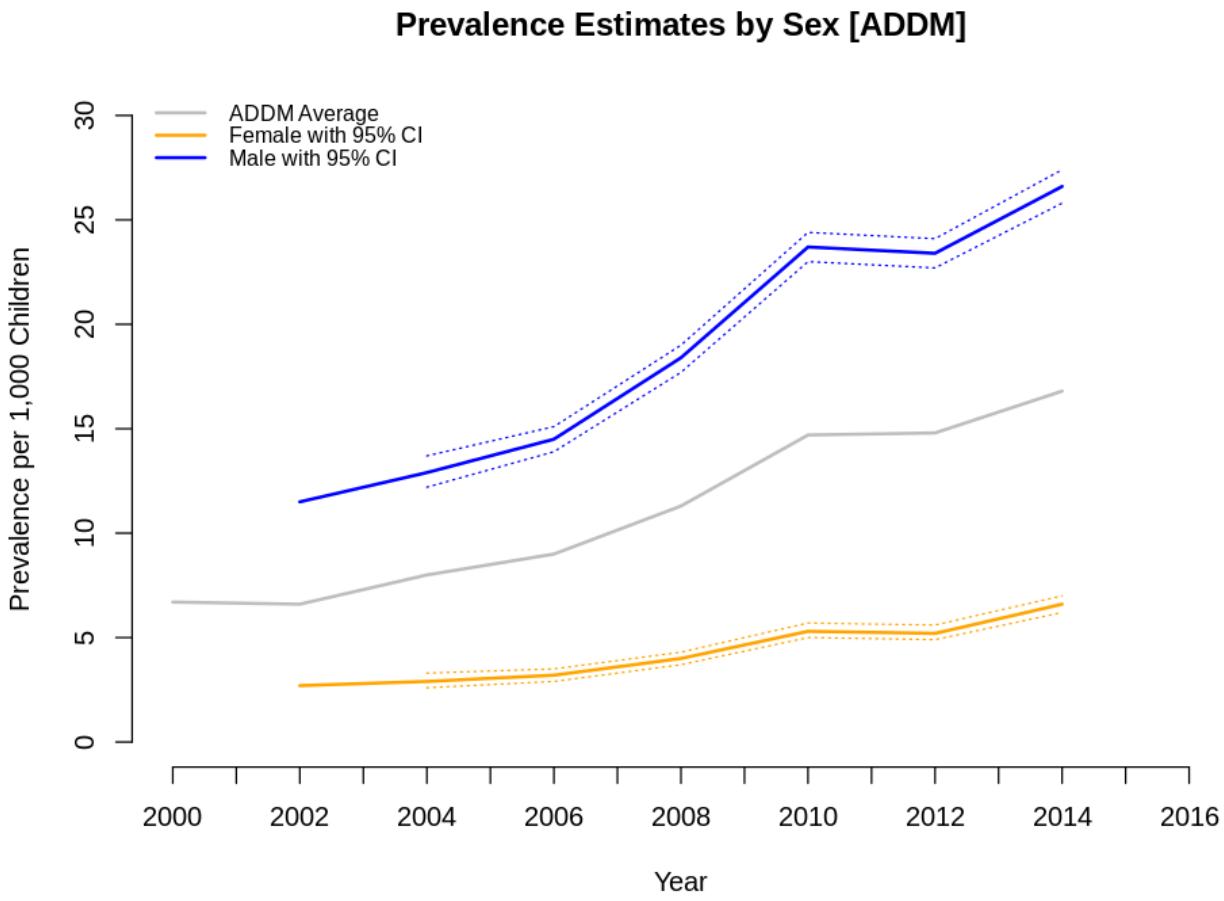
R pch: dot/point type: http://www.endmemo.com/program/R/pchsymbols.php (http://www.endmemo.com/program/R/pchsymbols.php)

R plot colour list: https://www.r-graph-gallery.com/42-colors-names.html (https://www.r-graph-gallery.com/42-colors-names.html)

Data Visualisation (Base Graphic) - [R] REPORTED PREVALENCE VARIES BY SEX [Source: ADDM] over [Year]

In [124]:

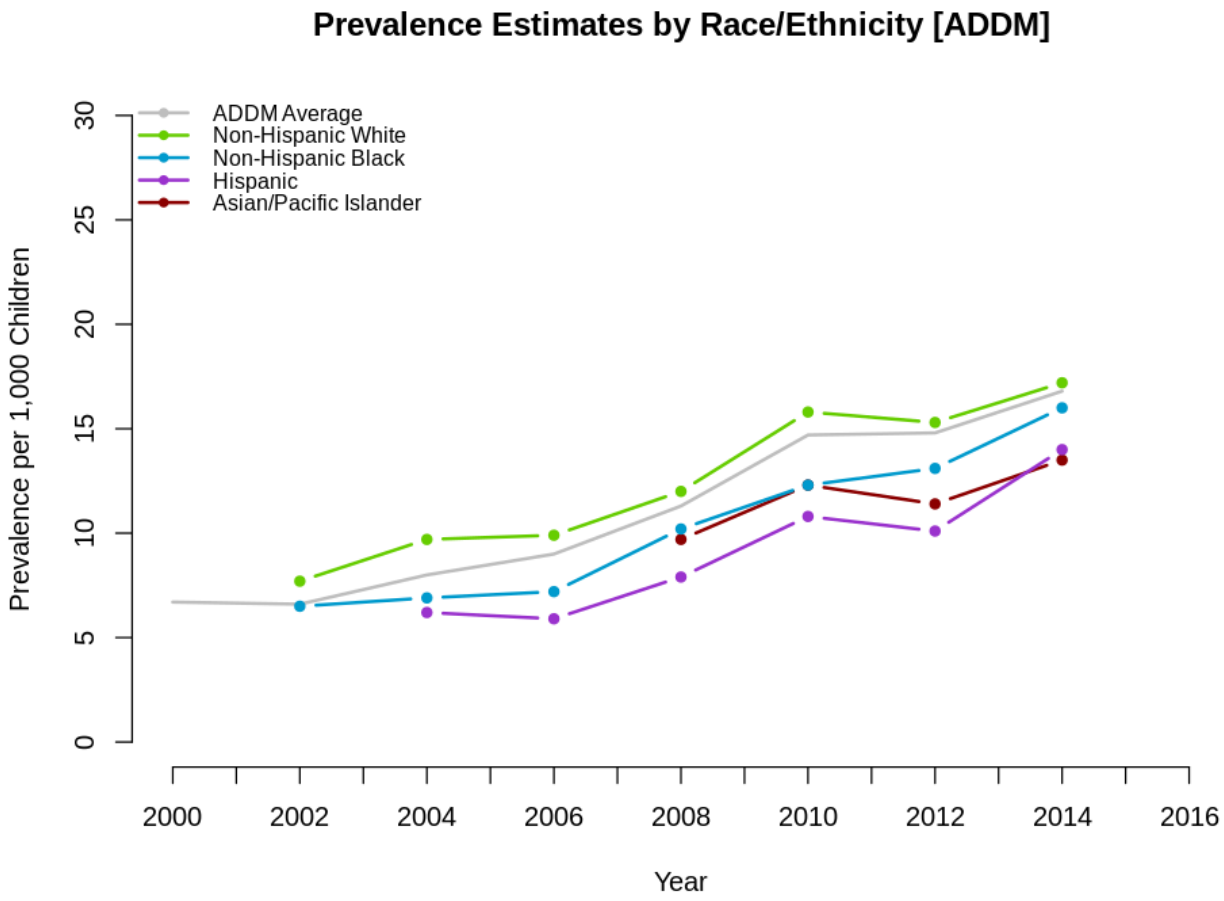
```
# ---------------------------------
# [addm] < Prevalence Varies by Sex >
# ---------------------------------
# Create a first line
plot(ASD_National$Year[ASD_National$Source == 'addm'],
     ASD_National$Prevalence[ASD_National$Source == 'addm'],
     col = "grey", lty = 1, lwd = 2,
     type = "l", # use dot/point
     pch = 0, # dot/point type: http://www.endmemo.com/program/R/pchsymbols.ph
     xlab="Year",
     xlim=c(2000, 2016), # Set x axis value range
     ylab="Prevalence per 1,000 Children",
     ylim=c(0, 30), # Set y axis value range
     main="Prevalence Estimates by Sex [ADDM]",
     col.main="black", col.lab="black", col.sub="grey",
     frame = FALSE, # Remove frame
     axes=FALSE # Remove x and y axis
)
axis(1, at=seq(2000, 2016, 1)) # Customize x axis
axis(2, at=seq(0, 30, 5)) # Customize y axis

# Add Female prevalence
lines(ASD_National$Year[ASD_National$Source == 'addm'],
      ASD_National$Female.Prevalence[ASD_National$Source == 'addm'],
      pch = 1, col = "orange", type = "l", lty = 1, lwd = 2)
# Add Female prevalence lower CI
lines(ASD_National$Year[ASD_National$Source == 'addm'],
      ASD_National$Female.Lower.CI[ASD_National$Source == 'addm'],
      pch = 1, col = "orange", type = "l", lty = 3, lwd = 1)
# Add Female prevalence upper CI
lines(ASD_National$Year[ASD_National$Source == 'addm'],
      ASD_National$Female.Upper.CI[ASD_National$Source == 'addm'],
      pch = 1, col = "orange", type = "l", lty = 3, lwd = 1)

# Add Male prevalence
lines(ASD_National$Year[ASD_National$Source == 'addm'],
      ASD_National$Male.Prevalence[ASD_National$Source == 'addm'],
      pch = 1, col = "blue", type = "l", lty = 1, lwd = 2)
# Add Male prevalence lower CI
lines(ASD_National$Year[ASD_National$Source == 'addm'],
      ASD_National$Male.Lower.CI[ASD_National$Source == 'addm'],
      pch = 1, col = "blue", type = "l", lty = 3, lwd = 1)
# Add Male prevalence upper CI
lines(ASD_National$Year[ASD_National$Source == 'addm'],
      ASD_National$Male.Upper.CI[ASD_National$Source == 'addm'],
      pch = 1, col = "blue", type = "l", lty = 3, lwd = 1)
# Add a legend to the plot
legend("topleft", legend=c('ADDM Average', 'Female with 95% CI', 'Male with 95
       col=c("grey", "orange", "blue"),
       #       pch = 20, # dot in a line
       lty = 1, # line type
       lwd = 2, # line width
       cex=0.8, # size of text
       bty = 'n' # Without frame
)
```



Data Visualisation (Base Graphic) - [R] REPORTED PREVALENCE VARIES BY RACE AND ETHNICITY [Source: ADDM]

In [125]:

```
# ---------------------------------
# [addm] < Prevalence Varies by Race and Ethnicity >
# ---------------------------------
# Create a first line
plot(ASD_National$Year[ASD_National$Source == 'addm'],
     ASD_National$Prevalence[ASD_National$Source == 'addm'],
     col = "grey", lty = 1, lwd = 2,
     type = "l", # use dot/point
     pch = 0, # dot/point type: http://www.endmemo.com/program/R/pchsymbols.ph
     xlab="Year",
     xlim=c(2000, 2016), # Set x axis value range
     ylab="Prevalence per 1,000 Children",
     ylim=c(0, 30), # Set y axis value range
     main="Prevalence Estimates by Race/Ethnicity [ADDM]",
     col.main="black", col.lab="black", col.sub="grey",
     frame = FALSE, # Remove frame
     axes=FALSE # Remove x and y axis
)
axis(1, at=seq(2000, 2016, 1)) # Customize x axis
axis(2, at=seq(0, 30, 5)) # Customize y axis

# R plot colour list: https://www.r-graph-gallery.com/42-colors-names.html

# Add Asian.or.Pacific.Islander.Prevalence
lines(ASD_National$Year[ASD_National$Source == 'addm'],
      ASD_National$Asian.or.Pacific.Islander.Prevalence[ASD_National$Source ==
      pch = 20, col = "darkred", type = "b", lty = 1, lwd = 2)
# Add Hispanic.Prevalence
lines(ASD_National$Year[ASD_National$Source == 'addm'],
      ASD_National$Hispanic.Prevalence[ASD_National$Source == 'addm'],
      pch = 20, col = "darkorchid3", type = "b", lty = 1, lwd = 2)
# Add Non.hispanic.black.Prevalence
lines(ASD_National$Year[ASD_National$Source == 'addm'],
      ASD_National$Non.hispanic.black.Prevalence[ASD_National$Source == 'addm'
      pch = 20, col = "deepskyblue3", type = "b", lty = 1, lwd = 2)
# Add Non.hispanic.white.Prevalence
lines(ASD_National$Year[ASD_National$Source == 'addm'],
      ASD_National$Non.hispanic.white.Prevalence[ASD_National$Source == 'addm'
      pch = 20, col = "chartreuse3", type = "b", lty = 1, lwd = 2)

# Add a legend to the plot
legend("topleft", legend=c('ADDM Average',
                           'Non-Hispanic White',
                           'Non-Hispanic Black',
                           'Hispanic',
                           'Asian/Pacific Islander'),
       col=c("grey", "chartreuse3", "deepskyblue3", "darkorchid3", "darkred"),
       pch = 20, # dot in a line
       lty = 1, # line type
       lwd = 2, # line width
       cex=0.8, # size of text
       bty = 'n' # Without frame
)
```



In [126]:
```
# Adjust in-line plot size to M x N
options(repr.plot.width=8, repr.plot.height=4)
```

Quiz:

Add 95% Confidence Interval to above plot

In [127]:
```
# Write your code below and press Shift+Enter to execute
```

Double-click **here** for the solution.

Quiz:

Use talbe() to count No. prevalence records for each Data Source. Then use barplot() to visualize.

In [128]:
```
# Write your code below and press Shift+Enter to execute
```

Double-click **here** for the solution.

Quiz:

Which Data Sources are available in which years?

In [129]:
```
# Write your code below and press Shift+Enter to execute
```

Double-click **here** for the solution.

Quiz:

Which Data Source has breakdown Prevalvence data by sex/gender?

In [130]:
```
# Write your code below and press Shift+Enter to execute
```

Double-click **here** for the solution.

Quiz:

Which Data Source has breakdown Prevalvence data by race and ethnicity?

In [131]:
```
# Write your code below and press Shift+Enter to execute
```

Double-click **here** for the solution.

Excellent! You have completed the workshop notebook!

Connect with the author:

This notebook was written by GU Zhan (Sam) (https://sg.linkedin.com/in/zhan-gu-27a82823).

Sam (https://www.iss.nus.edu.sg/about-us/staff/detail/201/GU_Zhan) is currently a lecturer in Institute of Systems Science (https://www.iss.nus.edu.sg/) in National University of Singapore (http://www.nus.edu.sg/). He devotes himself into pedagogy & andragogy, and is very passionate in inspiring next generation of artificial intelligence lovers and leaders.

()

Appendices

Interactive workshops: < Learning R inside R > using swirl() (in R/RStudio)

https://github.com/telescopeuser/S-SB-Workshop (https://github.com/telescopeuser/S-SB-Workshop)

Neural Network 101 using nnet()

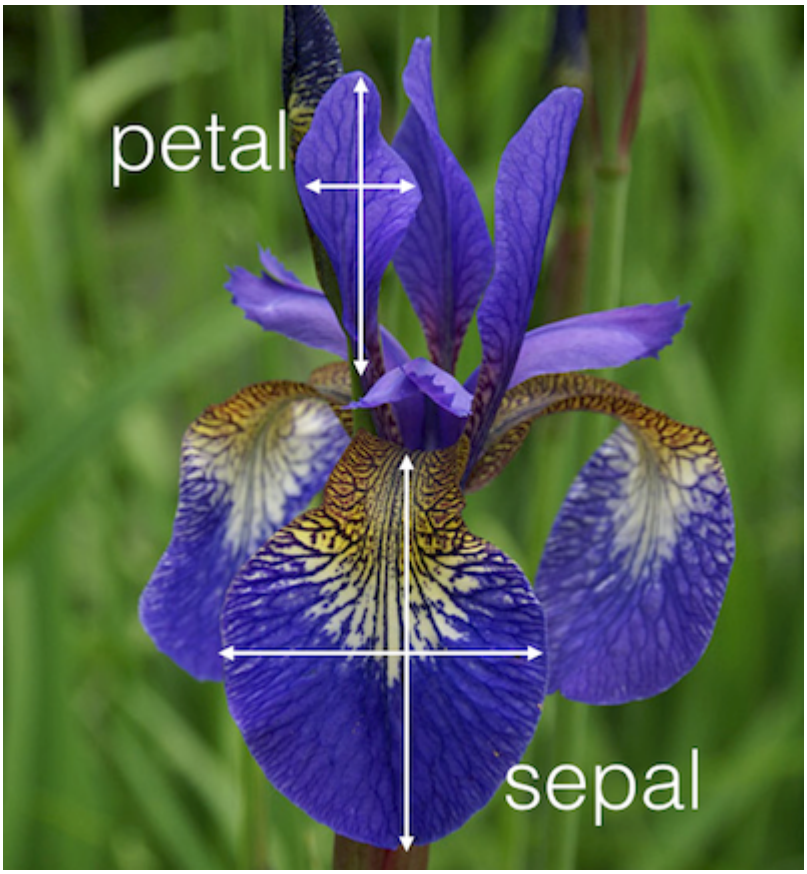
Use nerual net to classify three different species of iris flowers, based on four features/measurements of:

- length of the petals
- width of the petals
- length of the sepals
- width of the sepals

Iris setosa

Iris versicolor

Iris virginica



In [132]:
```r
# ---------------------------------
# Neural Network 101 using nnet()
# ---------------------------------
if(!require(nnet)){install.packages("nnet")}
library("nnet")
# ?nnet

# < Case: predict three different iris flower types >

# https://en.wikipedia.org/wiki/Iris_flower_data_set
# https://archive.ics.uci.edu/ml/datasets/iris

# Data preparation: split iris data in two halves, for training & testing resp
ir <- rbind(iris3[,,1],iris3[,,2],iris3[,,3])
targets <- class.ind( c(rep("setosa", 50), rep("versicolor", 50), rep("virgini
samp <- c(sample(1:50,25), sample(51:100,25), sample(101:150,25))
# Model training (machine learning / data fitting)
ir1 <- nnet(ir[samp,], targets[samp,], size = 2, rang = 0.1,
            decay = 5e-4, maxit = 200)
# Model evaluation function
test.cl <- function(true, pred) {
  true <- max.col(true)
  cres <- max.col(pred)
  table(true, cres)
}
# Model evaluation
test.cl(targets[-samp,], predict(ir1, ir[-samp,]))
```

```
Loading required package: nnet

# weights:  19
initial  value 55.502976
iter  10 value 30.733586
iter  20 value 25.156727
iter  30 value 25.119943
iter  40 value 23.843253
iter  50 value 19.077811
iter  60 value 18.385094
iter  70 value 18.157698
iter  80 value 18.066800
iter  90 value 18.038896
iter 100 value 18.005927
iter 110 value 18.003539
iter 120 value 17.998391
iter 130 value 17.984949
iter 140 value 17.467577
iter 150 value 14.523387
iter 160 value 14.241485
iter 170 value 13.370945
iter 180 value 3.910266
iter 190 value 3.298636
iter 200 value 3.115464
final  value 3.115464
stopped after 200 iterations

    cres
true  1  2  3
   1 25  0  0
   2  0 23  2
   3  0  0 25
```