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

One-Stop Analytics: Exploratory Data Analysis (EDA)

Case Study of Autism Spectrum Disorder (ASD) with R

[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.

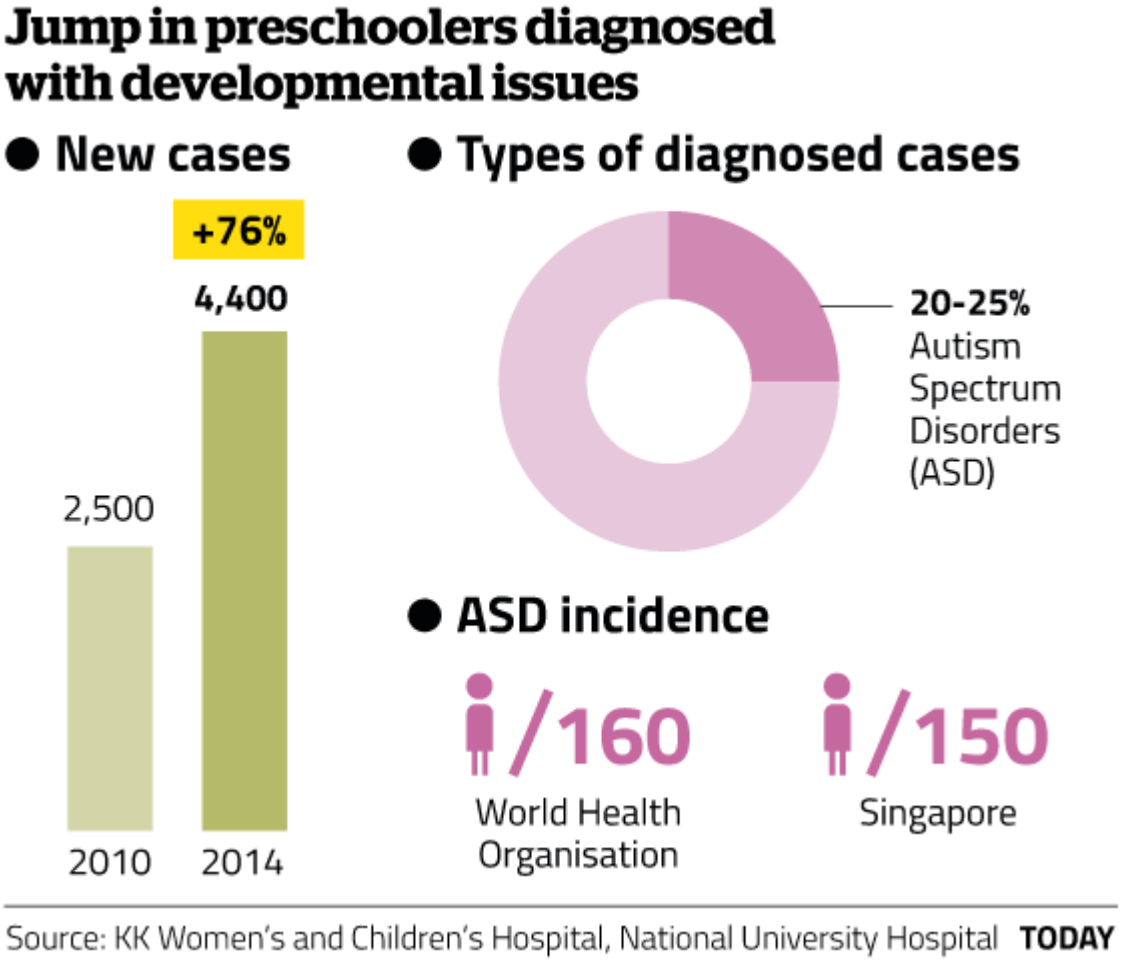
https://www.cdc.gov/ncbddd/autism/data/index.html (https://www.cdc.gov/ncbddd/autism/data/index.html)

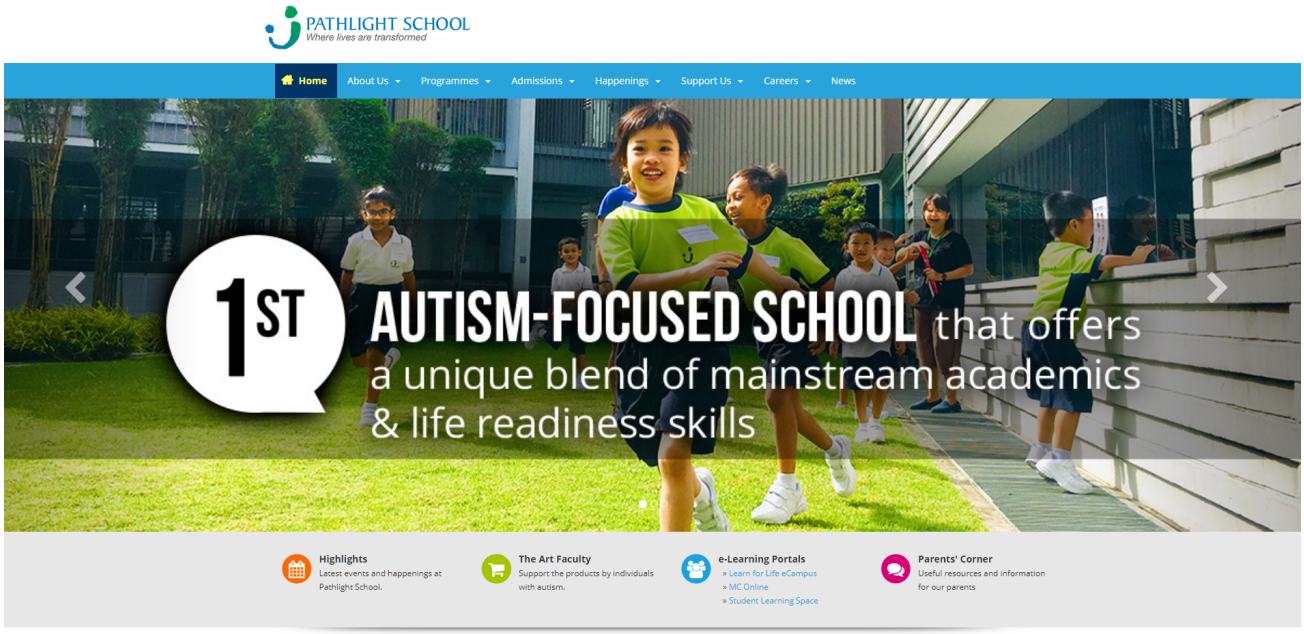
[Singapore]

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)

- **EDA - Summarization**
- **Data Visualisation (Enhanced)**
- **Workshop Submission**
- **Appendices**

()

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

EDA - Summarization

EDA - Summarization - High Level Data Summary

In [4]:
```
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
High:28           High     : 8      2004   : 4
Low :14           Low      :14      2008   : 4
                  Medium   :18      2012   : 4
                  Very High: 2      2000   : 3
                                    2002   : 3
                                    2006   : 3
                                    (Other):21
```

Data Visualisation (Enhanced)

In [5]:
```
if(!require(ggplot2)){install.packages("ggplot2")}
library(ggplot2)
```

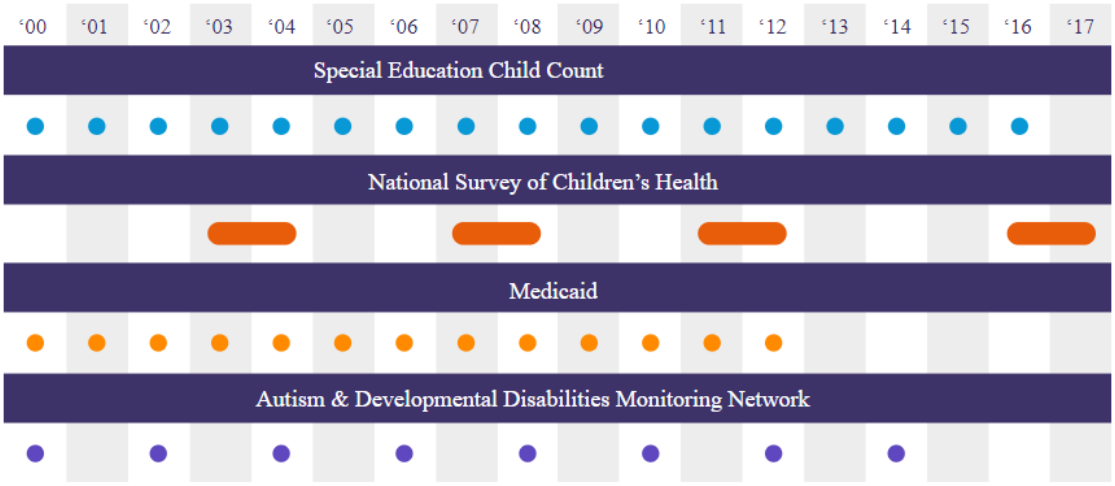
```
Loading required package: ggplot2
Registered S3 methods overwritten by 'ggplot2':
  method         from
  [.quosures     rlang
  c.quosures     rlang
  print.quosures rlang
```

In [6]:
```
# Adjust in-line plot size to M x N
options(repr.plot.width=8, repr.plot.height=4)
```

Data Visualisation (Enhanced) - [CDC] Explore the Data

Years Data Available

Select state: U.S. or Total†



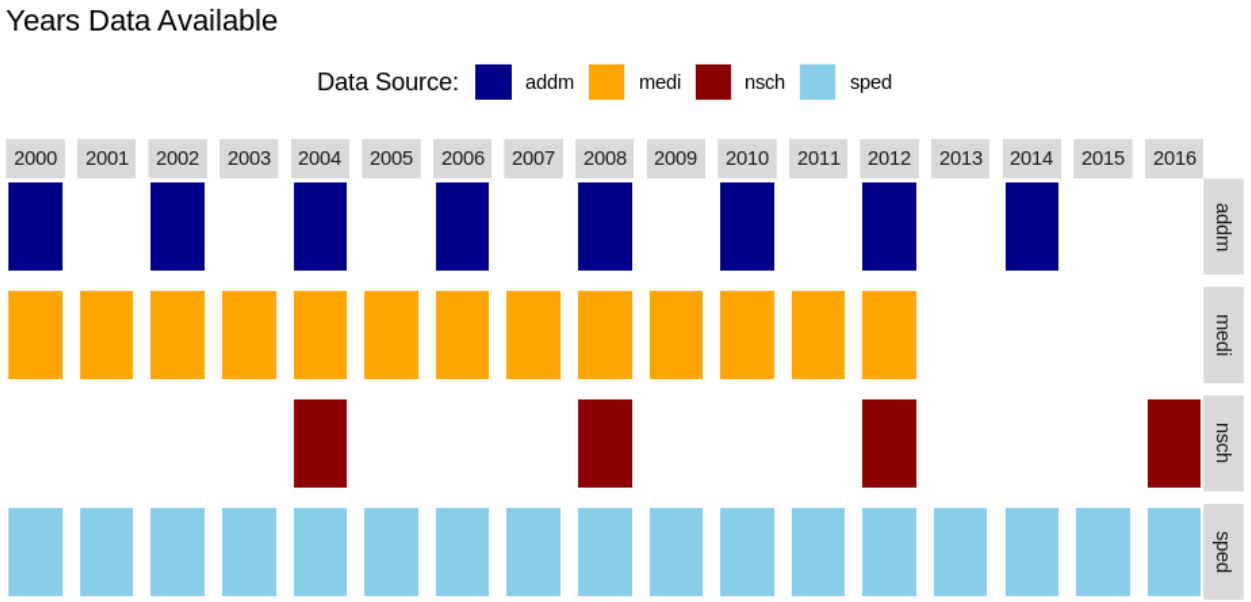
WHY THIS MATTERS

Because ASD data are collected at specific times, they provide a snapshot of what was going on at a certain moment in time. Findings from different data sources are typically reported a year or more *after* the data were collected; therefore, prevalence may have changed between the time data were collected and the time they were reported.

**ADDM estimate = the total for all sites combined.*

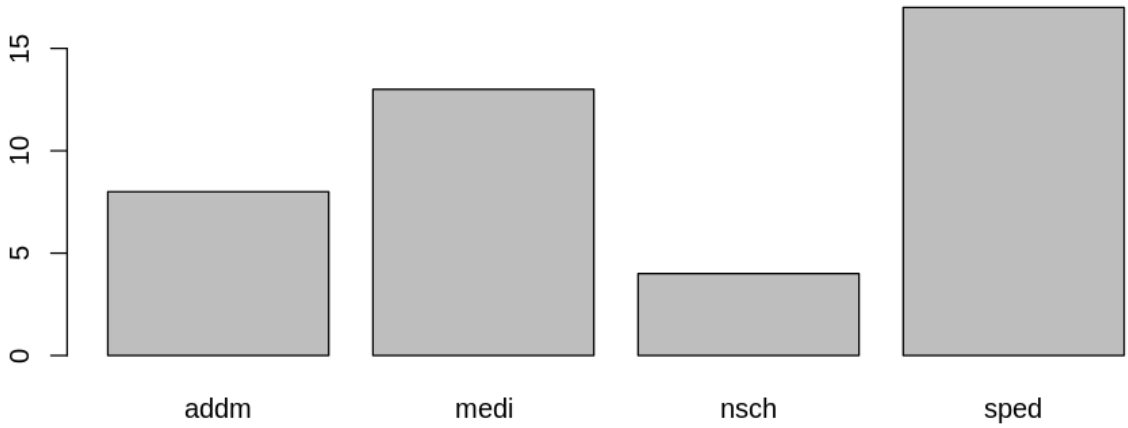
Data Visualisation (Enhanced) - [R] Explore the Data

In [7]:
```
# ---------------------------------
# [National] < Years Data Available >
# ---------------------------------
p = ggplot(ASD_National, aes(x = 1, fill = Source)) +
  geom_bar() + theme(axis.text.x=element_blank(),  # Hide axis
                     axis.ticks.x=element_blank(), # Hide axis
                     axis.text.y=element_blank(),  # Hide axis
                     axis.ticks.y=element_blank(), # Hide axis
                     panel.background = element_blank(), # Remove panel backgr
                     legend.position="top"
  ) +
  scale_fill_manual("Data Source:", values = c("addm" = "darkblue",
                                               "medi" = "orange",
                                               "nsch" = "darkred",
                                               "sped" = "skyblue")) +
  labs(x="", y="", title="Years Data Available") + # layers of graphics
  facet_grid(facets = Source~Year)
# Show plot
p
```

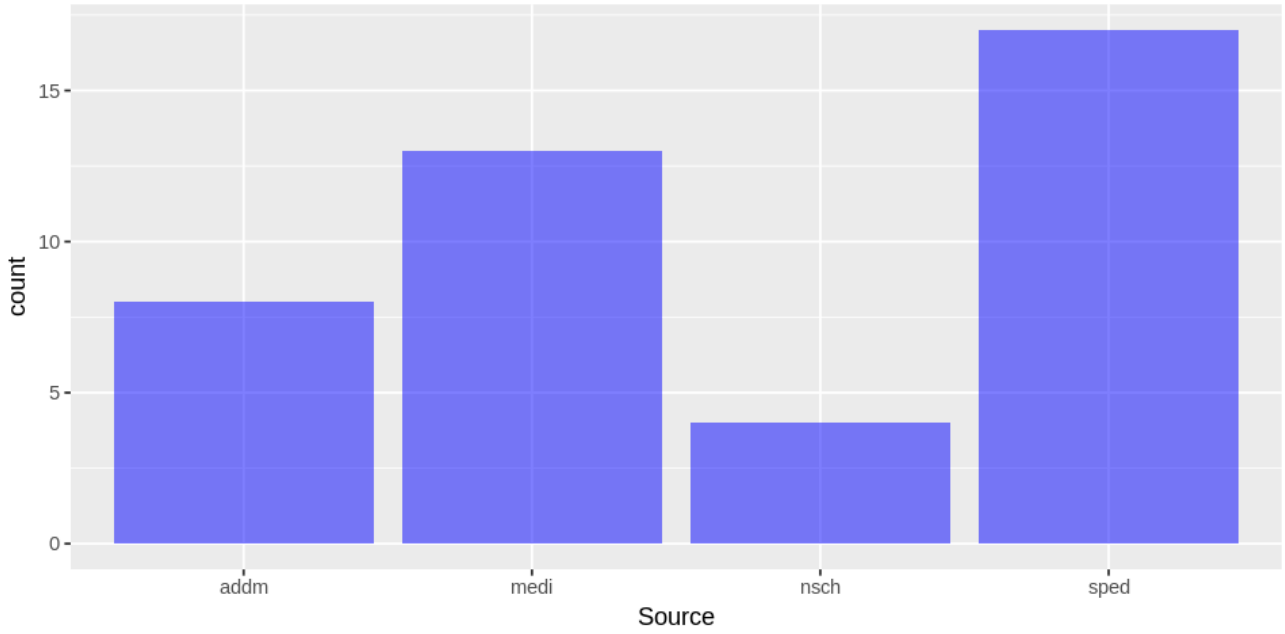


Data Visualisation (Enhanced) - Barplot

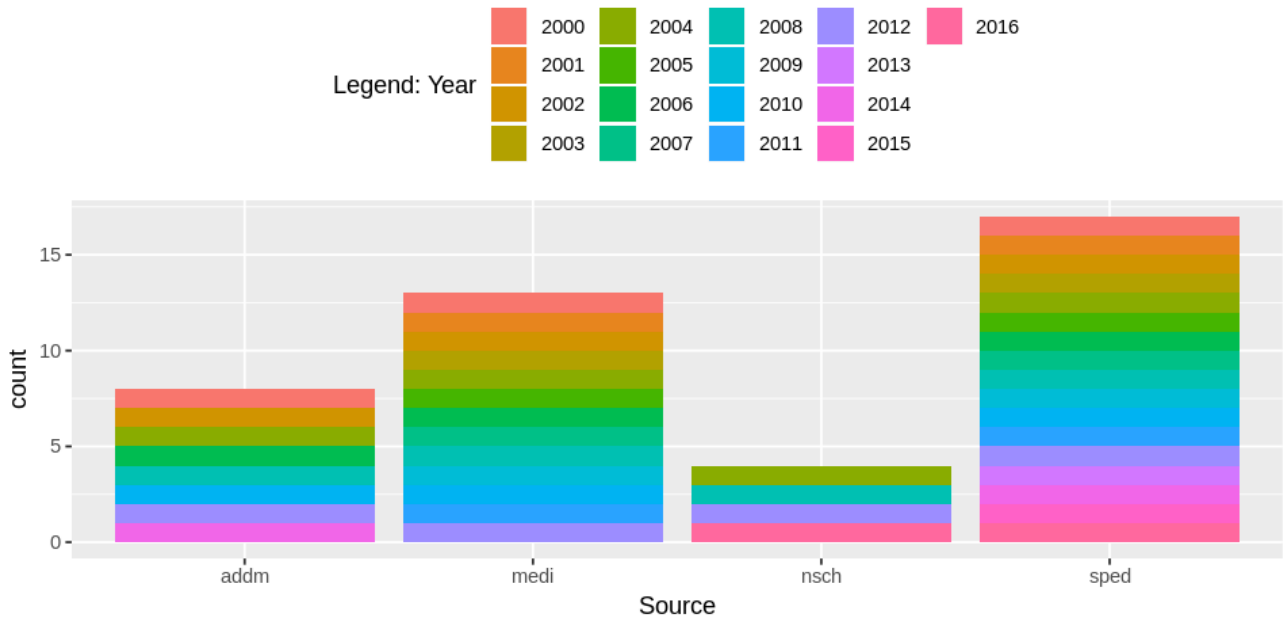
In [8]:
```
# Create bar chart using R graphics
barplot(table(ASD_National$Source))
```



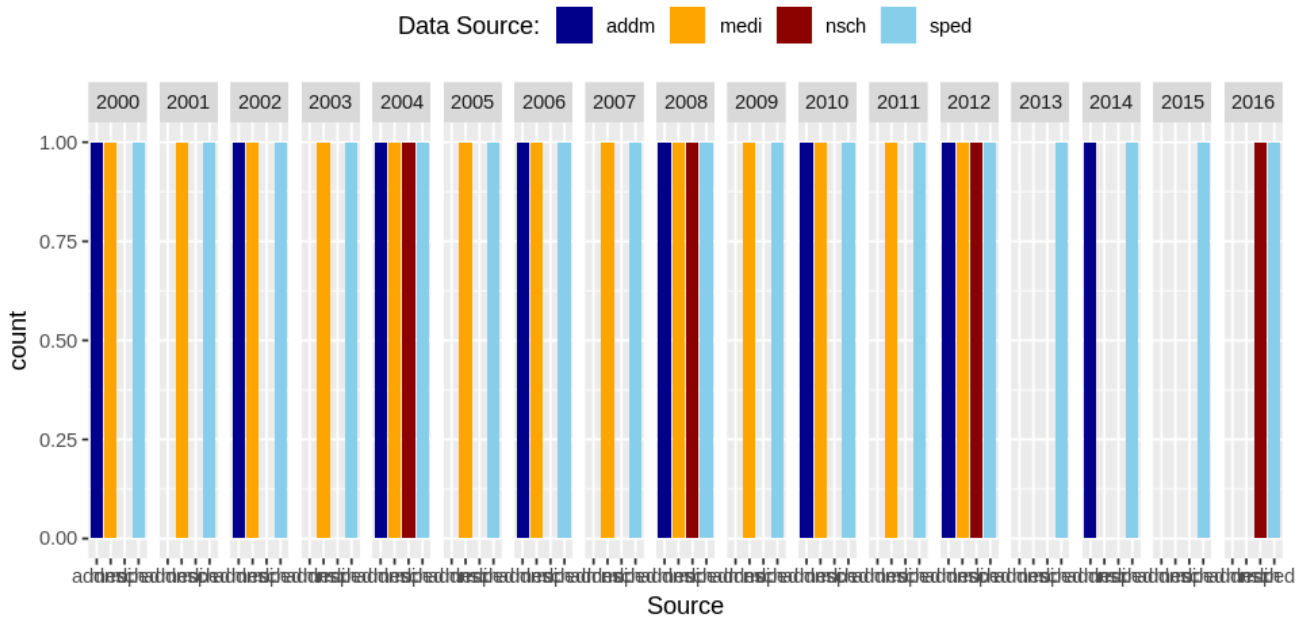
In [9]:
```
# Create bar chart using ggplot2
ggplot(ASD_National, aes(x = Source)) + geom_bar(fill = "blue", alpha=0.5)
```



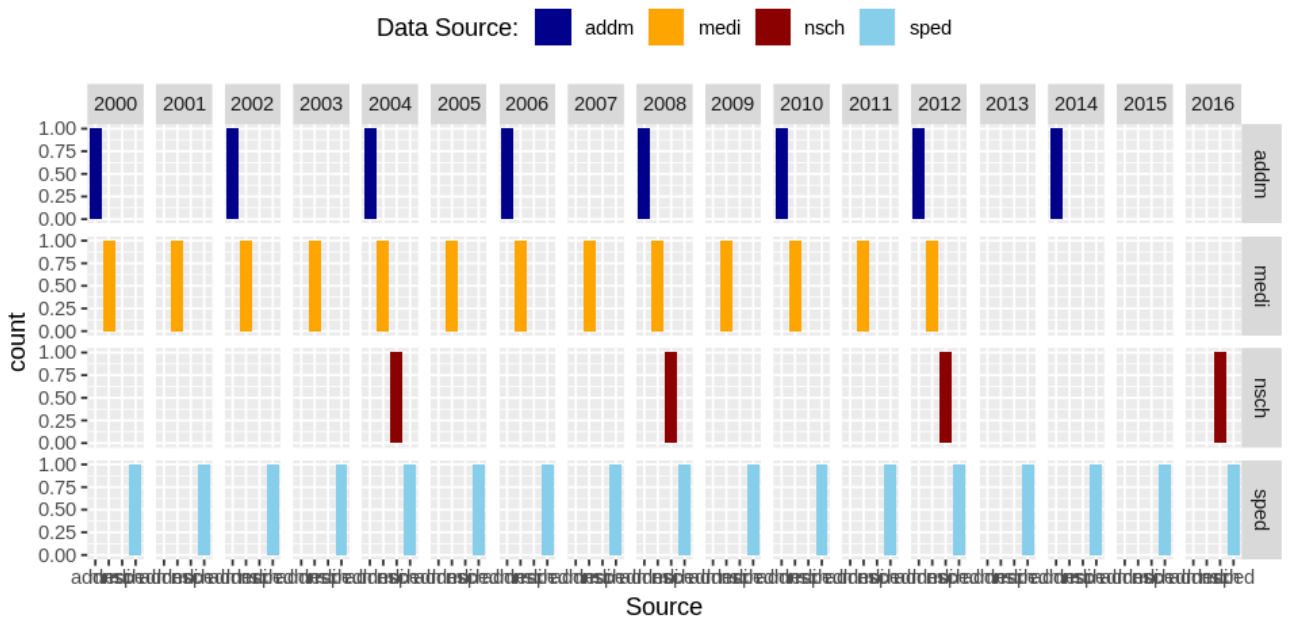
In [10]:
```
# Use color to differentiate sub-group data (Year)
ggplot(ASD_National, aes(x = Source, fill = factor(Year))) + geom_bar() +
  theme(legend.position="top") + labs(fill = "Legend: Year")
```



In [11]:
```
# Split chart to mutiple columns by using: facets = . ~ Year
ggplot(ASD_National, aes(x = Source, fill = Source)) + geom_bar() +
  theme(legend.position="top") +
  scale_fill_manual("Data Source:", values = c("addm" = "darkblue",
                                               "medi" = "orange",
                                               "nsch" = "darkred",
                                               "sped" = "skyblue")) +

  facet_grid(facets = . ~ Year)
```



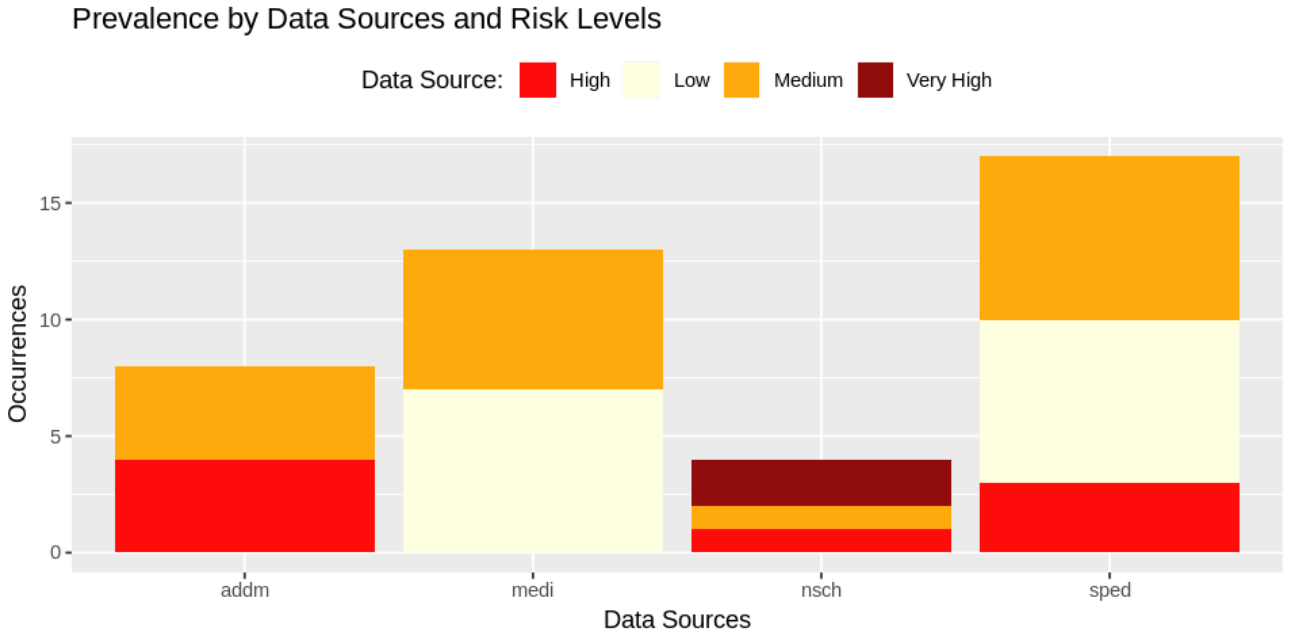
In [12]:
```
# Split chart to mutiple rows and columns by using: facets = Source ~ Year
ggplot(ASD_National, aes(x = Source, fill = Source)) + geom_bar() +
  theme(legend.position="top") +
  scale_fill_manual("Data Source:", values = c("addm" = "darkblue",
                                               "medi" = "orange",
                                               "nsch" = "darkred",
                                               "sped" = "skyblue")) +
  facet_grid(facets = Source~Year)
```



Above chart is now very similar to earlier [National] < Years Data Available >.

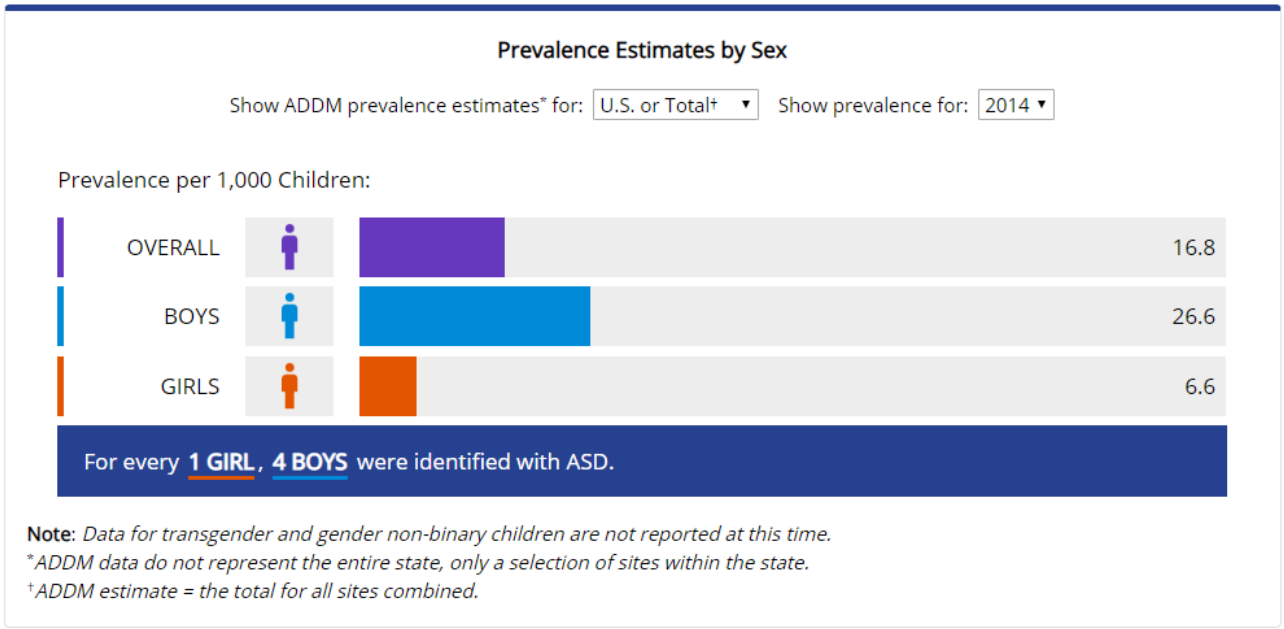
Data Visualisation (Enhanced) - [R] Prevalence by Data Sources and Risk Levels

In [13]:
```
# Use color to differentiate sub-group data (Year)
ggplot(ASD_National, aes(x = Source, fill = Prevalence_Risk4)) +
  geom_bar(alpha=0.95, position = position_stack(reverse = TRUE)) + # Reverse
  scale_fill_manual("Data Source:", values = c("Low" = "lightyellow",
                                               "Medium" = "orange",
                                               "High" = "red",
                                               "Very High" = "darkred")) +
  labs(x="Data Sources", y="Occurrences", title="Prevalence by Data Sources an
  theme(legend.position="top") + labs(fill = "Legend: Risk")
```



Barplot / Column plot

Data Visualisation (Enhanced) - [CDC] REPORTED PREVALENCE VARIES BY SEX



Data Visualisation (Enhanced) - [R] REPORTED PREVALENCE VARIES BY SEX [Source: ADDM] [Year: 2014]

In [14]:
```
# Filter only data of ADDM
ASD_National_ADDM <- subset(ASD_National, Source == 'addm')
#
ASD_National_ADDM
```

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	
addm	2002	6.6	6.8	6.3	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	11.5	
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
addm	2012	14.8	15.2	14.4	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	23.4	2
addm	2014	16.8	17.3	16.4	Autism & Developmental Disabilities Monitoring Network	addm-Autism & Developmental Disabilities Monitoring Network	26.6	2

In [15]:
```
# Construct a new re-shaped dataframe of [ Source: ADDM ] [Year: 2014]
#
Process_Source = 'addm'
Process_Year = 2014
```

Define a function to create a re-shaped dataframe:

In [16]:
```
Function_Reshape_ASD_National_ADDM <- function(Process_Source, Process_Year) {
    # Create the vectors:
    Sex.Group  = c('Overall',
                   'Boys',
                   'Girls')
    Sex.Group

    Prevalence = c(ASD_National_ADDM$Prevalence[ASD_National_ADDM$Year == Proc
                   ASD_National_ADDM$Male.Prevalence[ASD_National_ADDM$Year ==
                   ASD_National_ADDM$Female.Prevalence[ASD_National_ADDM$Year
    Prevalence

    # Combine all the vectors into a data frame:
    ASD_National_ADDM_Reshaped_DF = data.frame(Sex.Group, Prevalence, stringsA

    # Add new columns:
    ASD_National_ADDM_Reshaped_DF$Source = Process_Source
    ASD_National_ADDM_Reshaped_DF$Year = Process_Year
    return(ASD_National_ADDM_Reshaped_DF) # Return a dataframe
}
```

Use defined function Function_Reshape_ASD_National_ADDM( ) for a specific year:

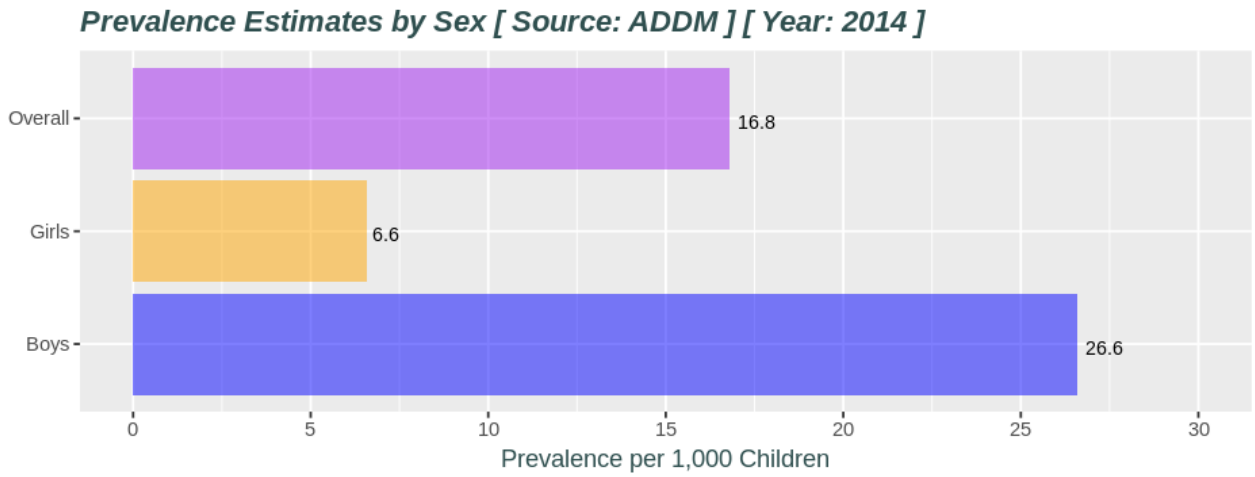
In [17]:
```
ASD_National_ADDM_Reshaped_DF <- Function_Reshape_ASD_National_ADDM(Process_So
ASD_National_ADDM_Reshaped_DF
```

Sex.Group	Prevalence	Source	Year
Overall	16.8	addm	2014
Boys	26.6	addm	2014
Girls	6.6	addm	2014

Visualise: **Prevalence Estimates by Sex [Source: ADDM] [Year: 2014]**

In [18]:
```
# Adjust in-line plot size to M x N
options(repr.plot.width=8, repr.plot.height=3)
```

In [19]:
```
ggplot(ASD_National_ADDM_Reshaped_DF, aes(Sex.Group, Prevalence)) +
  geom_col(aes(fill = Sex.Group, colours = ), alpha=0.5) + # Use column chart
  geom_text(aes(label = Prevalence), vjust = +0.75, hjust = -0.2, size = 3) +
  scale_y_continuous(name = "Prevalence per 1,000 Children",
                     breaks = seq(0, 30, 5),
                     limits=c(0, 30)) +
  scale_x_discrete(name = "") +
  scale_fill_manual("Sex Group:", values = c("Overall" = "purple",
                                             "Boys" = "blue",
                                             "Girls" = "orange")) +
  ggtitle("Prevalence Estimates by Sex [ Source: ADDM ] [ Year: 2014 ]") +
  theme(title = element_text(face = 'bold.italic', color = "darkslategrey"),
        axis.title = element_text(face = 'plain', color = "darkslategrey"),
        legend.position = 'none') +
  coord_flip()  # Rotate chart
#  facet_grid(facets = Year ~ .)
```



Data Visualisation (Enhanced) - [R] REPORTED PREVALENCE VARIES BY SEX [Source: ADDM] [Year: ALL]

In [20]:
```
# Create a new datafarme to hold re-shaped data for all years.
ASD_National_ADDM_Reshaped_DF_All = ASD_National_ADDM_Reshaped_DF # Loaded wit
```

In [21]:
```
Process_Source = 'addm'
unique(ASD_National_ADDM$Year)
```

2000 2002 2004 2006 2008 2010 2012 2014

Use defined function Function_Reshape_ASD_National_ADDM( ) for ALL remaining years:

In [22]:
```
ASD_National_ADDM_Reshaped_DF <- Function_Reshape_ASD_National_ADDM(Process_Sc
ASD_National_ADDM_Reshaped_DF
# Append rows to existing dataframe, using Row Bind function: rbind()
ASD_National_ADDM_Reshaped_DF_All = rbind(ASD_National_ADDM_Reshaped_DF_All, A
```

Sex.Group	Prevalence	Source	Year
Overall	14.8	addm	2012
Boys	23.4	addm	2012
Girls	5.2	addm	2012

In [23]:
```
ASD_National_ADDM_Reshaped_DF <- Function_Reshape_ASD_National_ADDM(Process_Sc
ASD_National_ADDM_Reshaped_DF
# Append rows to existing dataframe, using Row Bind function: rbind()
ASD_National_ADDM_Reshaped_DF_All = rbind(ASD_National_ADDM_Reshaped_DF_All, A
```

Sex.Group	Prevalence	Source	Year
Overall	14.7	addm	2010
Boys	23.7	addm	2010
Girls	5.3	addm	2010

In [24]:
```
ASD_National_ADDM_Reshaped_DF <- Function_Reshape_ASD_National_ADDM(Process_Sc
ASD_National_ADDM_Reshaped_DF
# Append rows to existing dataframe, using Row Bind function: rbind()
ASD_National_ADDM_Reshaped_DF_All = rbind(ASD_National_ADDM_Reshaped_DF_All, A
```

Sex.Group	Prevalence	Source	Year
Overall	11.3	addm	2008
Boys	18.4	addm	2008
Girls	4.0	addm	2008

In [25]:
```
ASD_National_ADDM_Reshaped_DF <- Function_Reshape_ASD_National_ADDM(Process_Sc
ASD_National_ADDM_Reshaped_DF
# Append rows to existing dataframe, using Row Bind function: rbind()
ASD_National_ADDM_Reshaped_DF_All = rbind(ASD_National_ADDM_Reshaped_DF_All, A
```

Sex.Group	Prevalence	Source	Year
Overall	9.0	addm	2006
Boys	14.5	addm	2006
Girls	3.2	addm	2006

In [26]:
```
ASD_National_ADDM_Reshaped_DF <- Function_Reshape_ASD_National_ADDM(Process_Sc
ASD_National_ADDM_Reshaped_DF
# Append rows to existing dataframe, using Row Bind function: rbind()
ASD_National_ADDM_Reshaped_DF_All = rbind(ASD_National_ADDM_Reshaped_DF_All, A
```

Sex.Group	Prevalence	Source	Year
Overall	8.0	addm	2004
Boys	12.9	addm	2004
Girls	2.9	addm	2004

In [27]:
```
ASD_National_ADDM_Reshaped_DF <- Function_Reshape_ASD_National_ADDM(Process_Sc
ASD_National_ADDM_Reshaped_DF
# Append rows to existing dataframe, using Row Bind function: rbind()
ASD_National_ADDM_Reshaped_DF_All = rbind(ASD_National_ADDM_Reshaped_DF_All, A
```

Sex.Group	Prevalence	Source	Year
Overall	6.6	addm	2002
Boys	11.5	addm	2002
Girls	2.7	addm	2002

In [28]:
```
ASD_National_ADDM_Reshaped_DF <- Function_Reshape_ASD_National_ADDM(Process_Sc
ASD_National_ADDM_Reshaped_DF
# Append rows to existing dataframe, using Row Bind function: rbind()
ASD_National_ADDM_Reshaped_DF_All = rbind(ASD_National_ADDM_Reshaped_DF_All, A
```

Sex.Group	Prevalence	Source	Year
Overall	6.7	addm	2000
Boys	NA	addm	2000
Girls	NA	addm	2000

In [29]:
```
# Re-shaped ADDM data for ALL years:
ASD_National_ADDM_Reshaped_DF_All
```

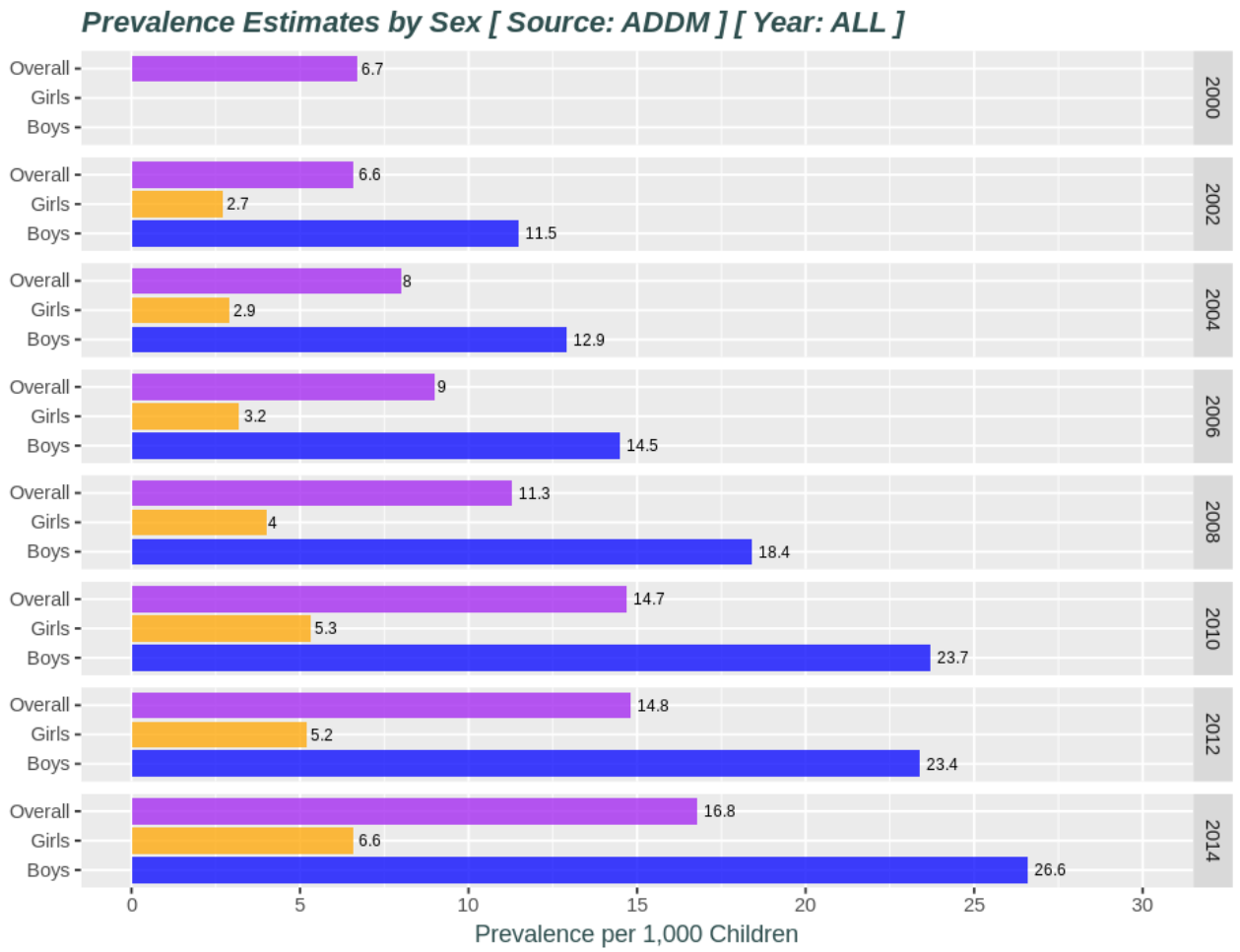
Sex.Group	Prevalence	Source	Year
Overall	16.8	addm	2014
Boys	26.6	addm	2014
Girls	6.6	addm	2014
Overall	14.8	addm	2012
Boys	23.4	addm	2012
Girls	5.2	addm	2012
Overall	14.7	addm	2010
Boys	23.7	addm	2010
Girls	5.3	addm	2010
Overall	11.3	addm	2008
Boys	18.4	addm	2008
Girls	4.0	addm	2008
Overall	9.0	addm	2006
Boys	14.5	addm	2006
Girls	3.2	addm	2006
Overall	8.0	addm	2004
Boys	12.9	addm	2004
Girls	2.9	addm	2004
Overall	6.6	addm	2002
Boys	11.5	addm	2002
Girls	2.7	addm	2002
Overall	6.7	addm	2000
Boys	NA	addm	2000
Girls	NA	addm	2000

Visualise: **Prevalence Estimates by Sex [Source: ADDM] [Year: ALL]**

In [30]:
```
# Adjust in-line plot size to M x N
options(repr.plot.width=8, repr.plot.height=6)
```

In [31]:
```
ggplot(ASD_National_ADDM_Reshaped_DF_All, aes(Sex.Group, Prevalence)) +
  geom_col(aes(fill = Sex.Group, colours = ), alpha=0.75) + # Use column chart
  geom_text(aes(label = Prevalence), vjust = +0.5, hjust = -0.2, size = 2.5) +
  scale_y_continuous(name = "Prevalence per 1,000 Children",
                     breaks = seq(0, 30, 5),
                     limits=c(0, 30)) +
  scale_x_discrete(name = "") +
  scale_fill_manual("Sex Group:", values = c("Overall" = "purple",
                                             "Boys" = "blue",
                                             "Girls" = "orange")) +
  ggtitle("Prevalence Estimates by Sex [ Source: ADDM ] [ Year: ALL ]") +
  theme(title = element_text(face = 'bold.italic', color = "darkslategrey"),
        axis.title = element_text(face = 'plain', color = "darkslategrey"),
        legend.position = 'none') +
  coord_flip() + # Rotate chart
  facet_grid(facets = Year ~ .)
```

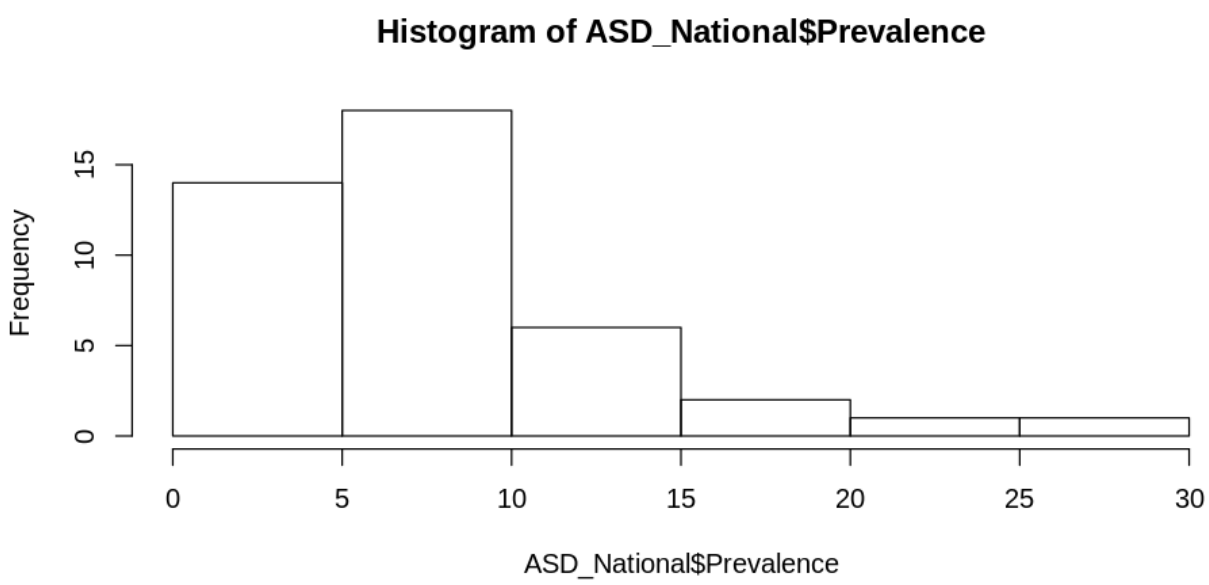
```
Warning message:
“Removed 2 rows containing missing values (position_stack).”Warning message:
“Removed 2 rows containing missing values (geom_text).”
```



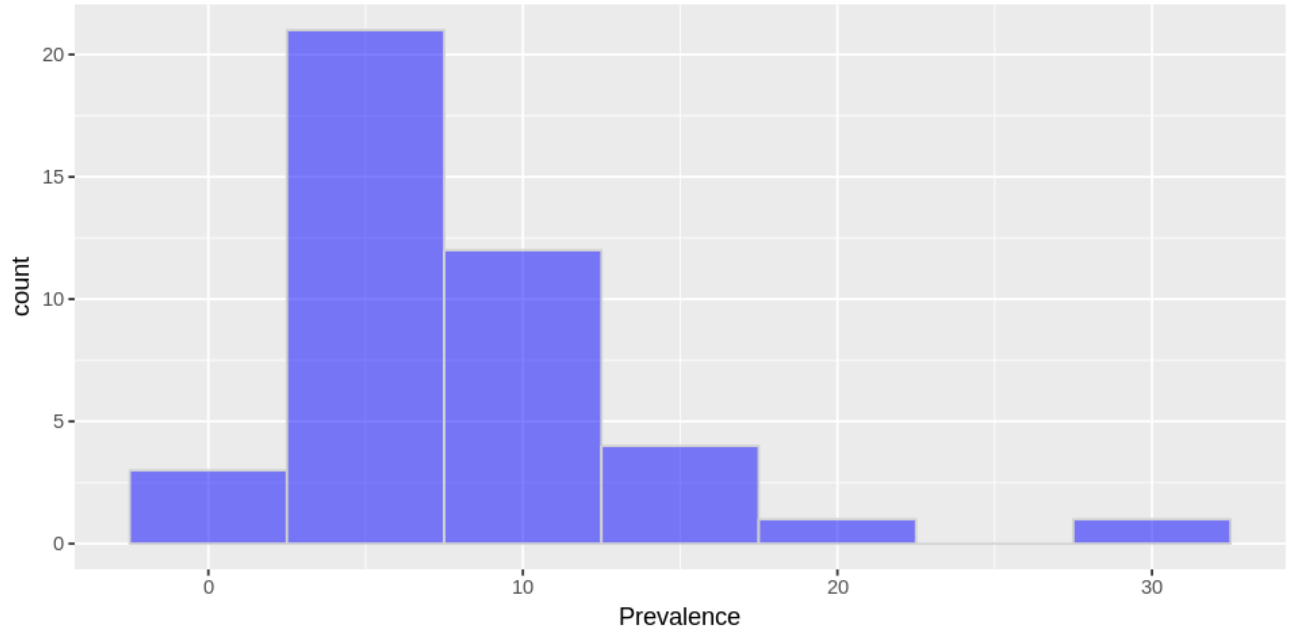
Data Visualisation (Enhanced) - Histogram (distribution of binned continuous variable)

In [32]:
```
# Adjust in-line plot size to M x N
options(repr.plot.width=8, repr.plot.height=4)
```

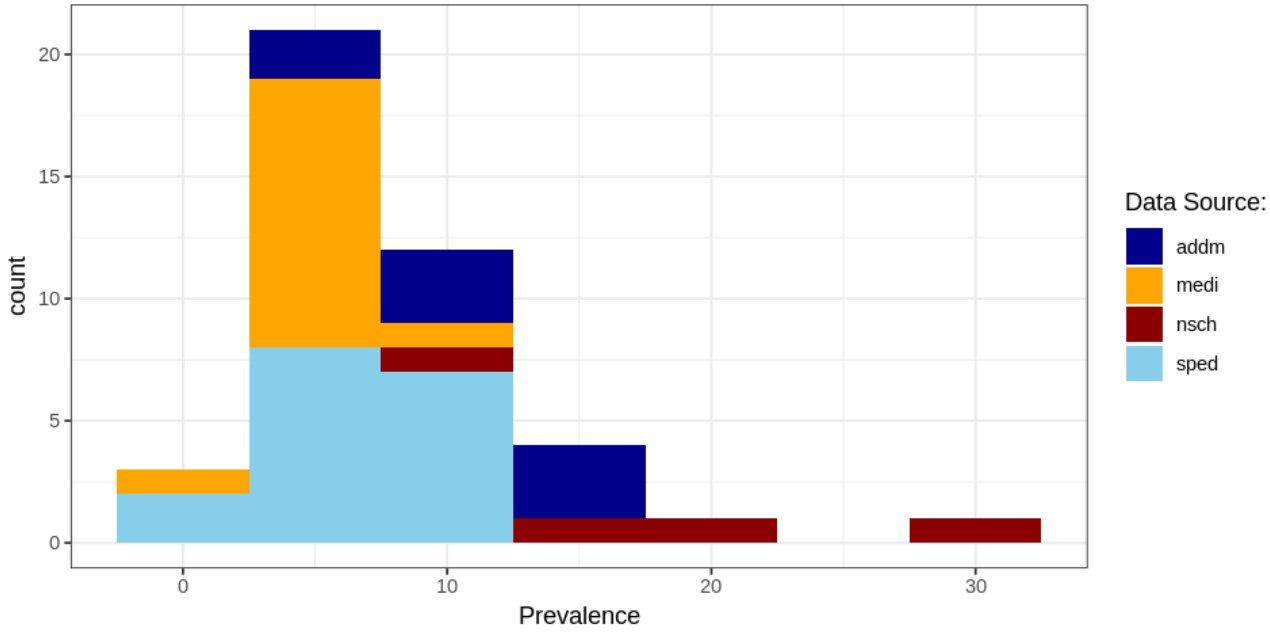
In [33]:
```
# Create histogram using R graphics
hist(ASD_National$Prevalence)
```



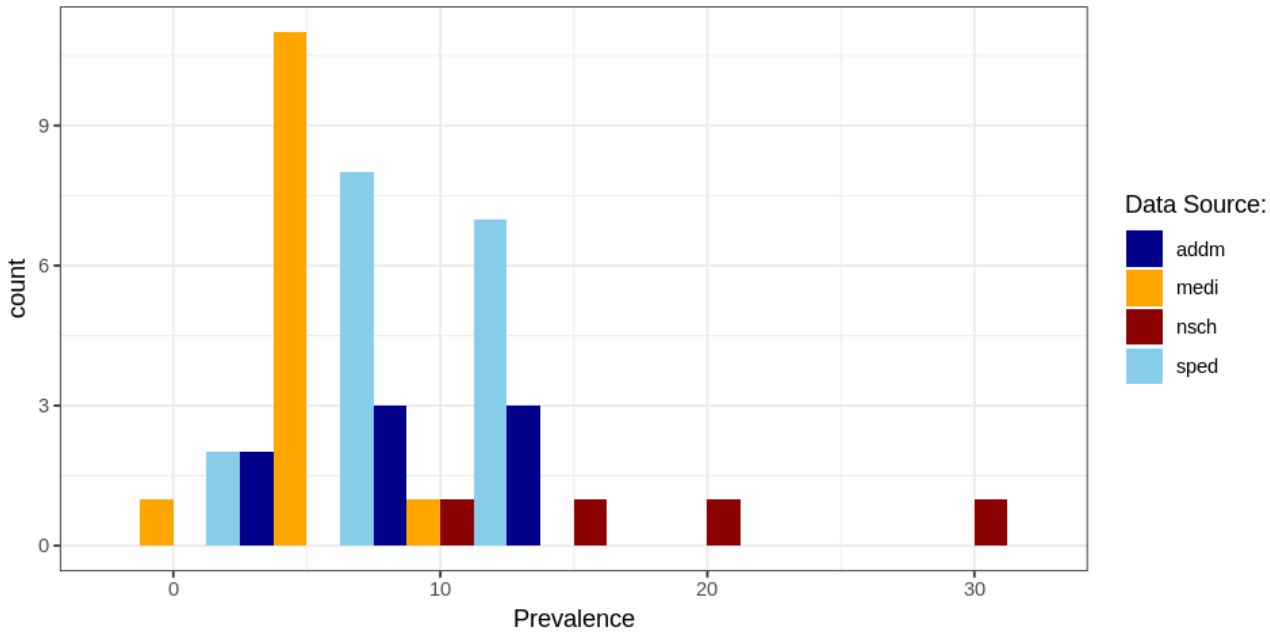
In [34]:
```
# Create histogram using ggplot2
ggplot(ASD_National, aes(x=Prevalence)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "lightgrey", alpha=0.5)
```



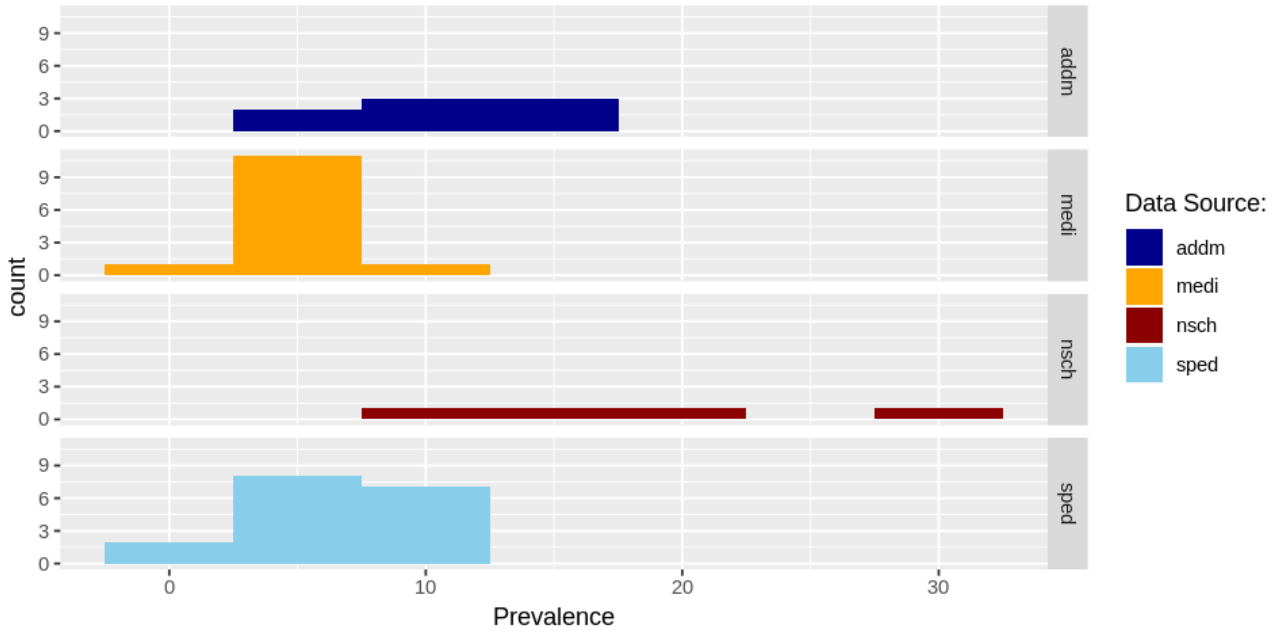
In [35]:
```
# Use color to differentiate sub-group data (Data Source)
ggplot(ASD_National, aes(x=Prevalence, fill = Source)) +
  geom_histogram(binwidth = 5) +
  theme_bw() + theme(legend.position="right") +
  scale_fill_manual("Data Source:", values = c("addm" = "darkblue",
                                               "medi" = "orange",
                                               "nsch" = "darkred",
                                               "sped" = "skyblue"))
```



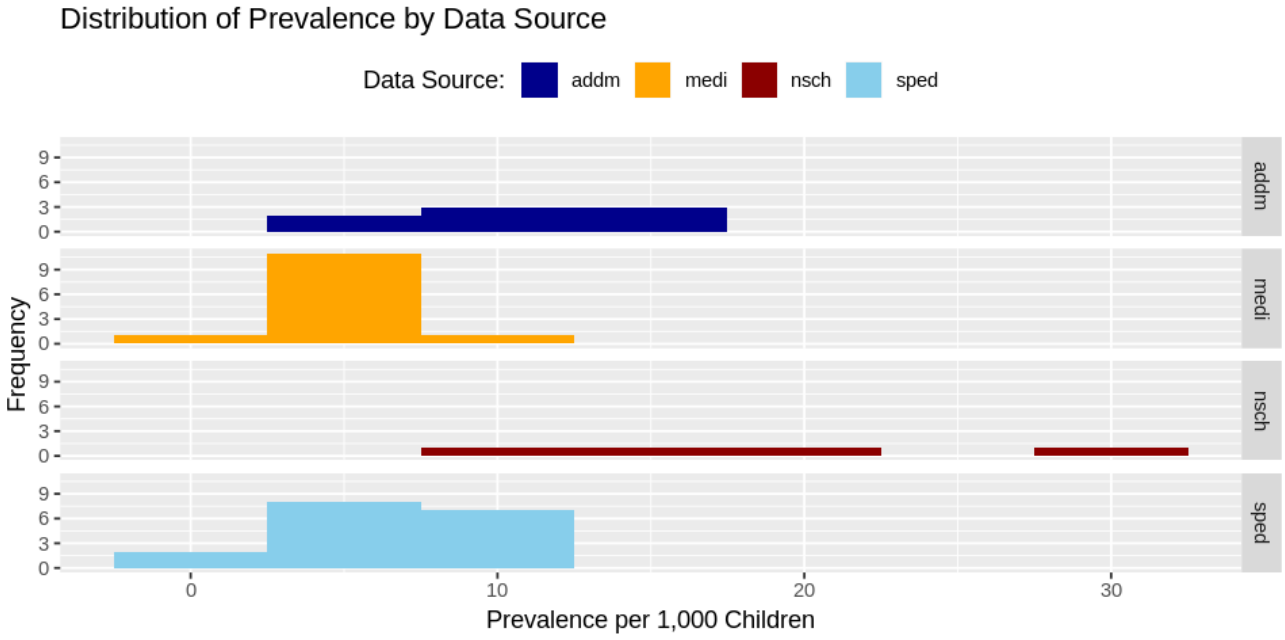
In [36]:
```
# Plot sub-group data side by side, using position="dodge"
ggplot(ASD_National, aes(x=Prevalence, fill = Source)) +
  geom_histogram(binwidth = 5, position="dodge") +
  theme_bw() + theme(legend.position="right") +
  scale_fill_manual("Data Source:", values = c("addm" = "darkblue",
                                               "medi" = "orange",
                                               "nsch" = "darkred",
                                               "sped" = "skyblue"))
```



In [37]:
```
# Split plots using facet_grid()
ggplot(ASD_National, aes(x=Prevalence, fill = Source)) +
  geom_histogram(binwidth = 5) +
  theme(legend.position="right") +
  scale_fill_manual("Data Source:", values = c("addm" = "darkblue",
                                               "medi" = "orange",
                                               "nsch" = "darkred",
                                               "sped" = "skyblue")) +
  facet_grid(facets = Source ~ .)
```



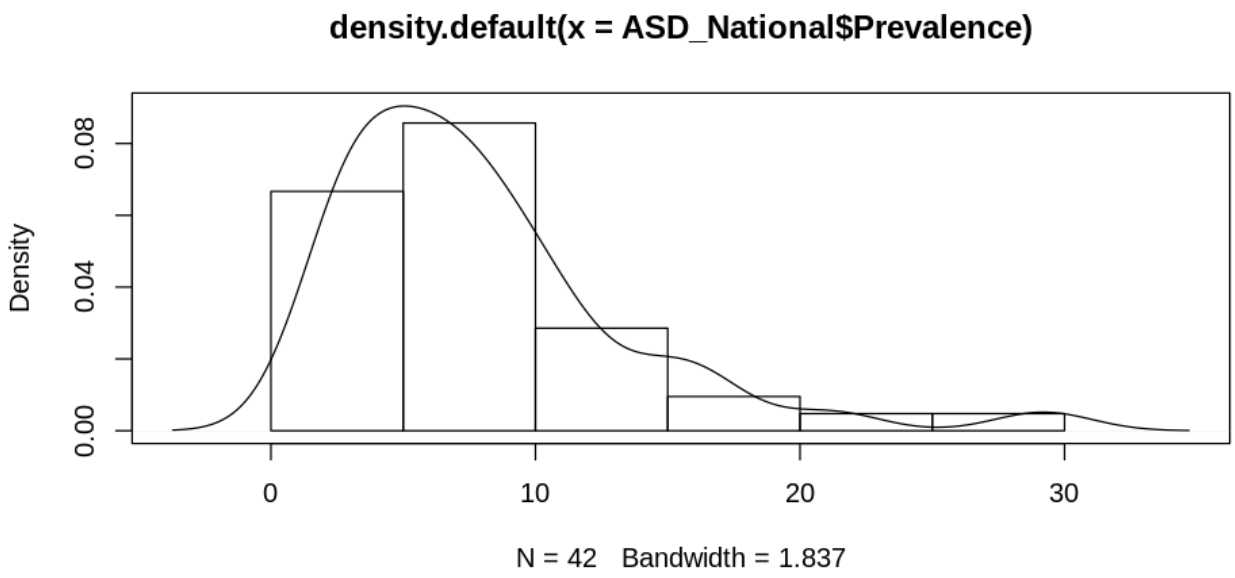
In [38]:
```
# Add title and caption using ggplot2
ggplot(ASD_National, aes(x=Prevalence, fill = Source)) +
  geom_histogram(binwidth = 5) +
  theme(legend.position="top") +
  scale_fill_manual("Data Source:", values = c("addm" = "darkblue",
                                               "medi" = "orange",
                                               "nsch" = "darkred",
                                               "sped" = "skyblue")) +
  labs(x="Prevalence per 1,000 Children",
       y="Frequency",
       title="Distribution of Prevalence by Data Source") +
  facet_grid(facets = Source ~ .)
```



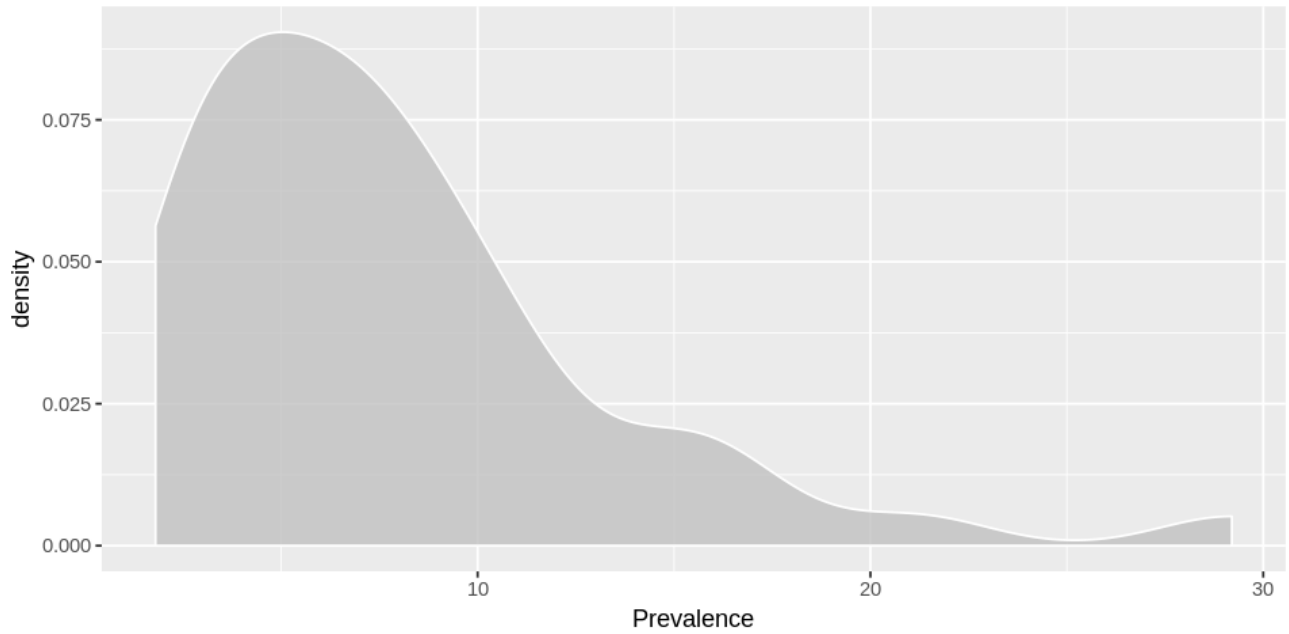
Data Visualisation (Enhanced) - Density plot (distribution for continuous variable normalized to 100% area under curve)

In [39]:
```
# Adjust in-line plot size to M x N
# options(repr.plot.width=8, repr.plot.height=4)
```

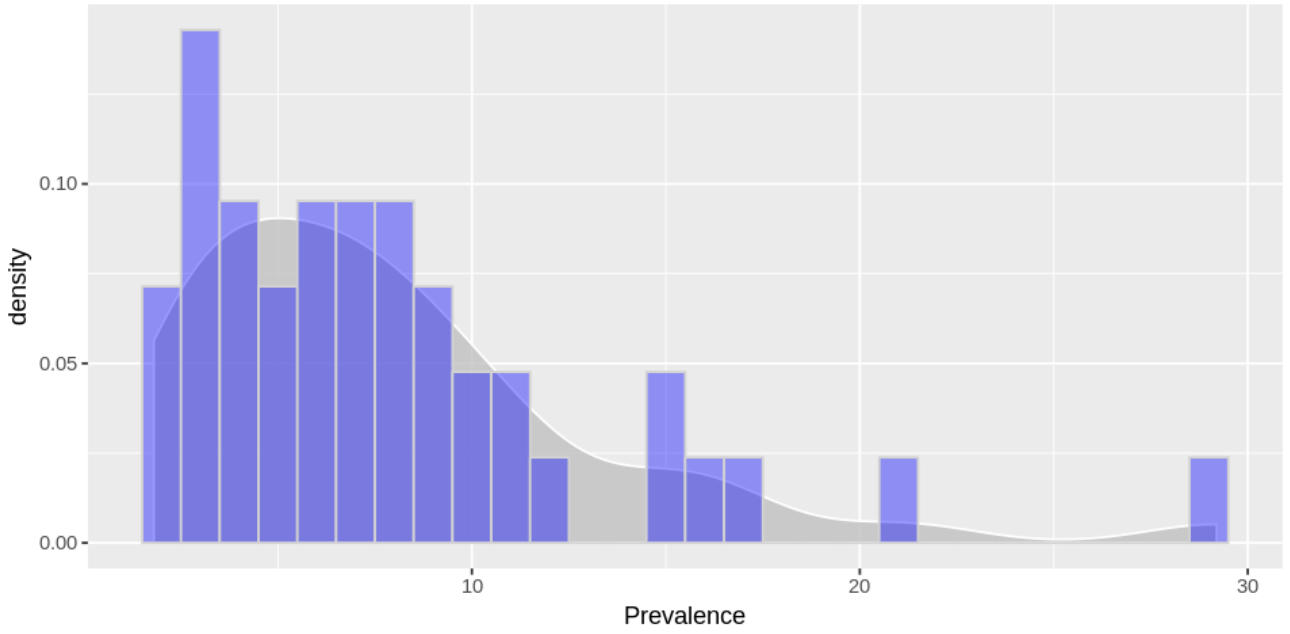
In [40]:
```
# Create plot using R graphics
plot(density(ASD_National$Prevalence))
# Optionally, overlay histogram
hist(ASD_National$Prevalence, probability = TRUE, add = TRUE)
```



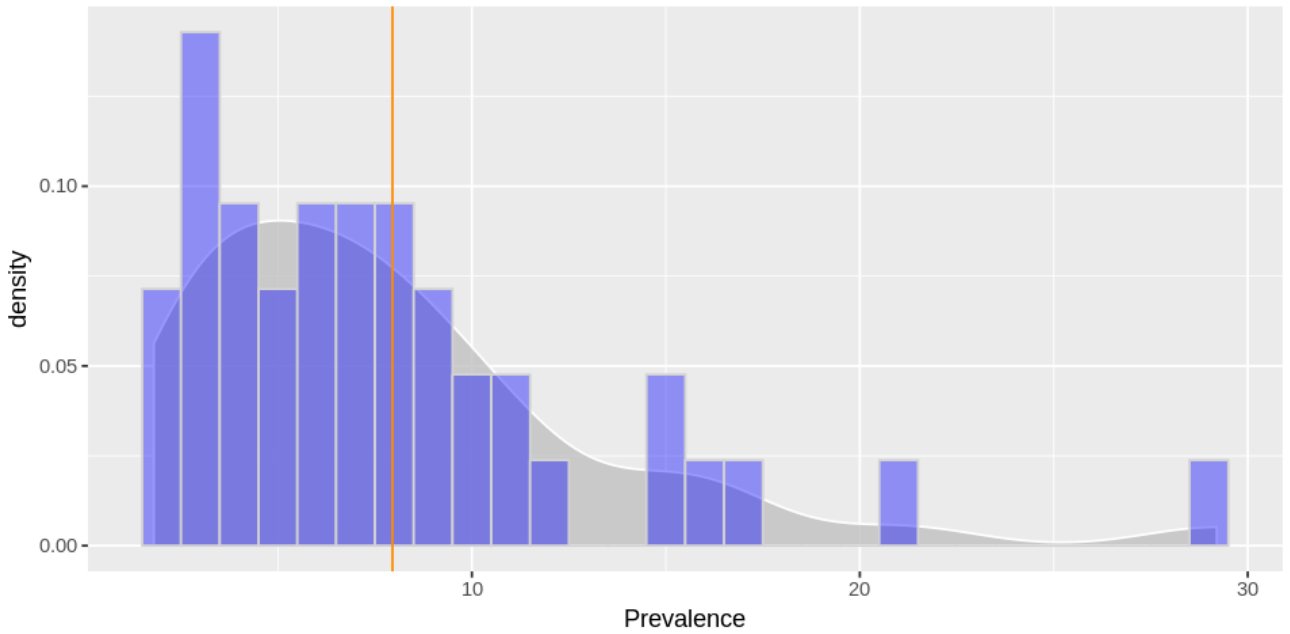
In [41]:
```
# Create plot using ggplot2
p <- ggplot(ASD_National) +
  geom_density(aes(x=Prevalence), fill = "grey", color = "white", alpha=0.75)
p # Show
```



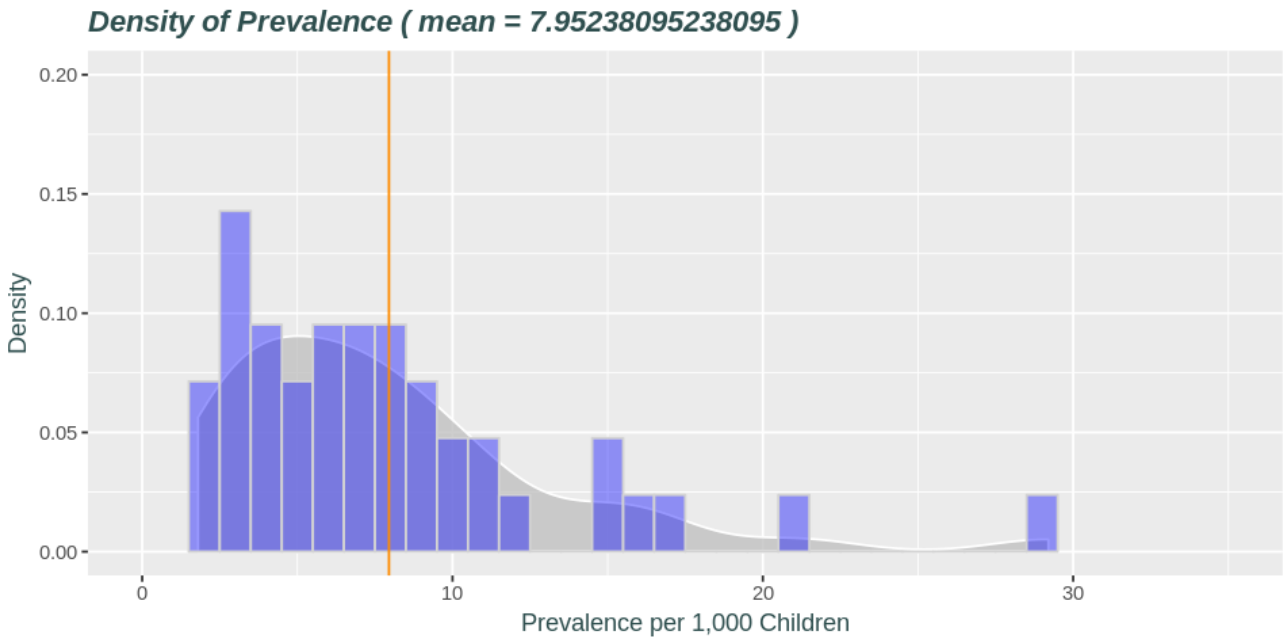
In [42]:
```
# Optionally, overlay histogram
p <- p + geom_histogram(aes(x = Prevalence, y = ..density..), binwidth = 1, fi
p # Show
```



In [43]:
```
# Optionally, overlay Prevalence mean
p <- p + geom_vline(aes(xintercept = mean(ASD_National$Prevalence)), colour="d
p # Show
```

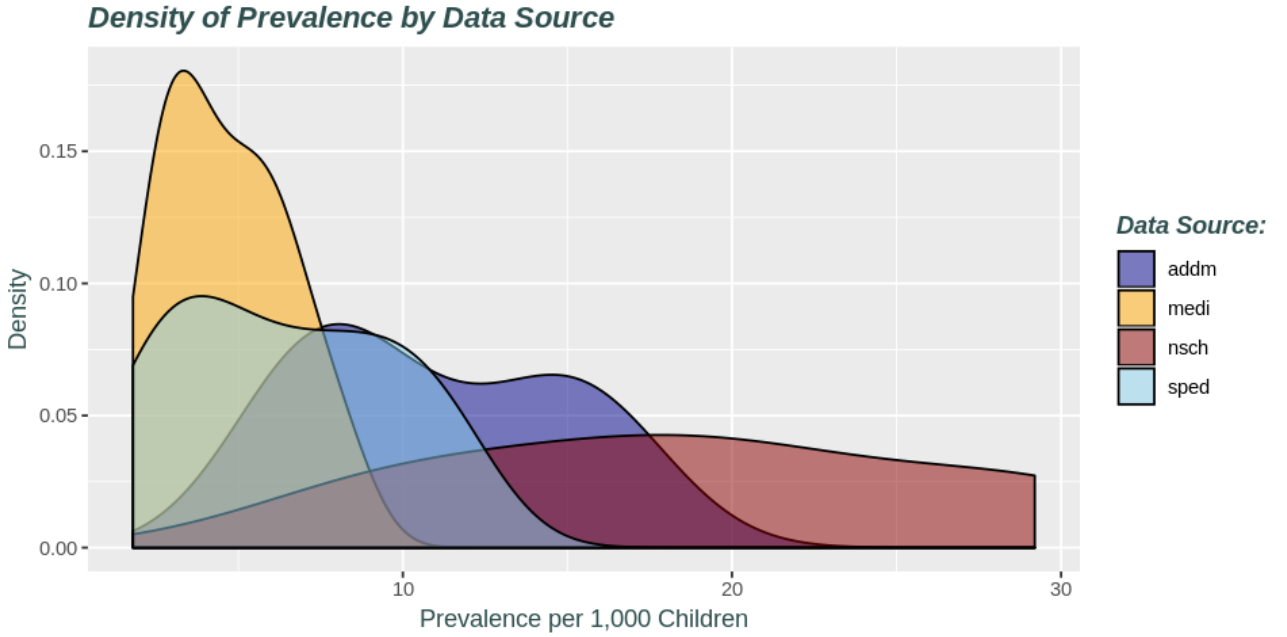


In [44]:
```
# Lastly, add other captions
p <- p + coord_cartesian(xlim=c(0, 35), ylim=c(0, 0.2)) +
  labs(x="Prevalence per 1,000 Children", y="Density",
       title=paste("Density of Prevalence ( mean =", mean(ASD_National$Prevale
  theme(title = element_text(face = 'bold.italic', color = "darkslategrey"),
        axis.title = element_text(face = 'plain', color = "darkslategrey"))
p # Show
```



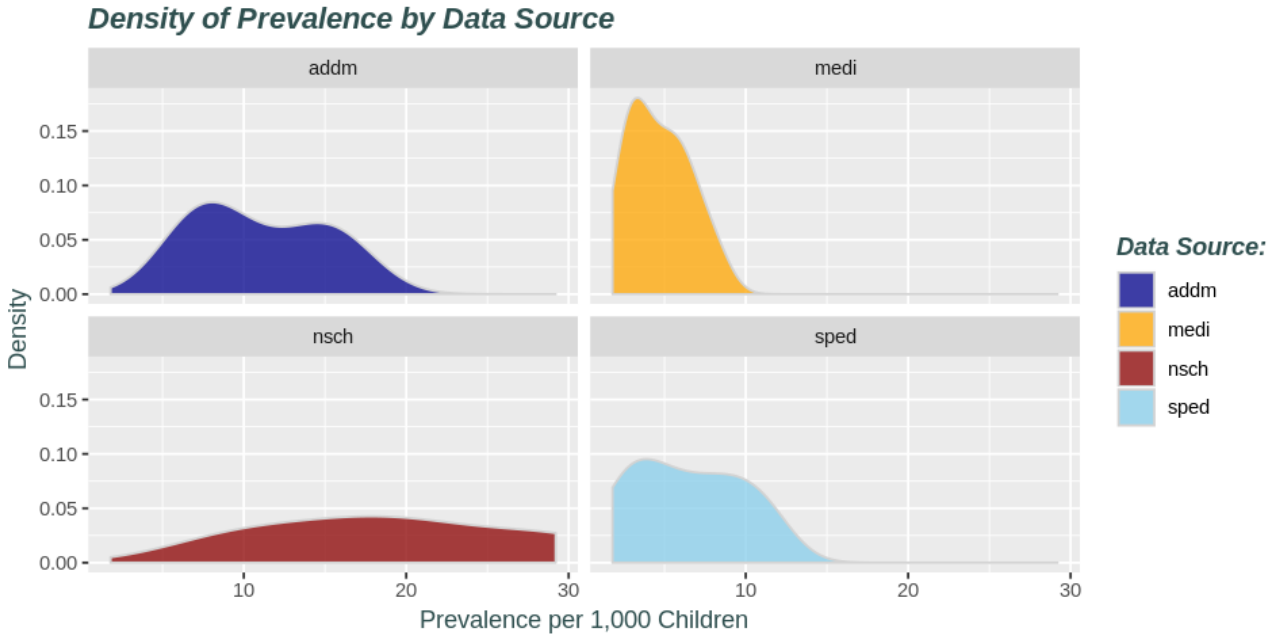
< Prevelance distribution by Data Source >

In [45]:
```
# Prevelance distribution by Data Source
ggplot(ASD_National) + geom_density(aes(x = Prevalence, fill = Source), alpha
  scale_fill_manual("Data Source:", values = c("addm" = "darkblue",
                                               "medi" = "orange",
                                               "nsch" = "darkred",
                                               "sped" = "skyblue")) +
  labs(x="Prevalence per 1,000 Children",
       y="Density",
       title="Density of Prevalence by Data Source") +
  theme(title = element_text(face = 'bold.italic', color = "darkslategrey"),
        axis.title = element_text(face = 'plain', color = "darkslategrey"))
```



< Prevelance distribution by Data Source with split >

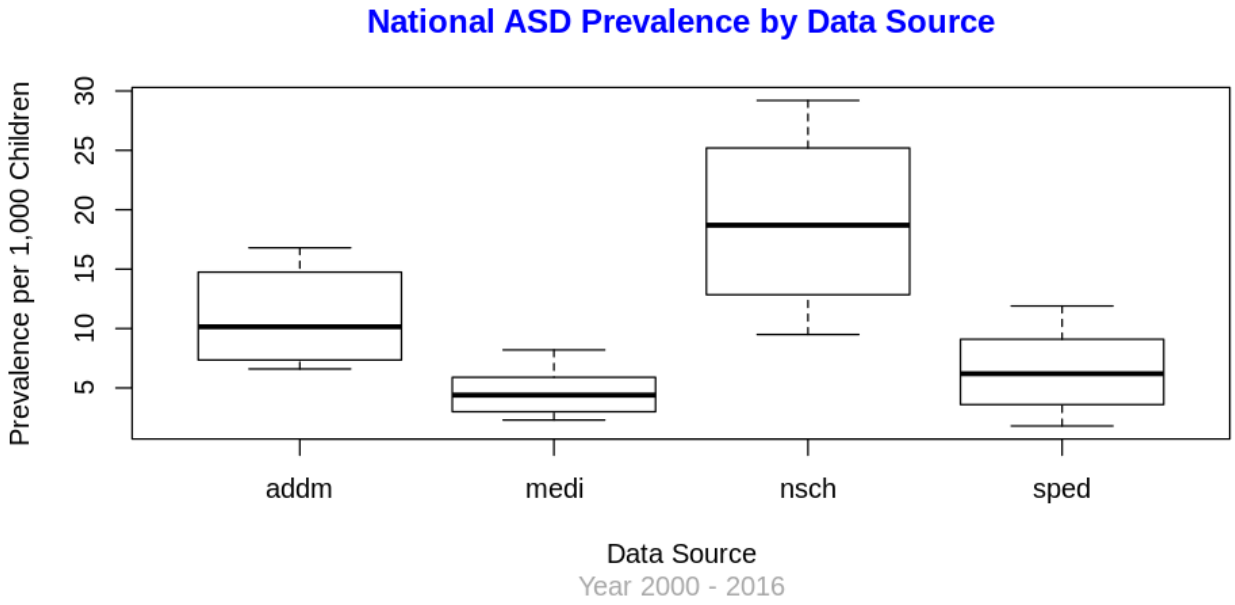
In [46]:
```
# Prevelance distribution by Data Source with split
ggplot(ASD_National) + geom_density(aes(x = Prevalence, fill = Source), colour
  scale_fill_manual("Data Source:", values = c("addm" = "darkblue",
                                               "medi" = "orange",
                                               "nsch" = "darkred",
                                               "sped" = "skyblue")) +
  labs(x="Prevalence per 1,000 Children",
       y="Density",
       title="Density of Prevalence by Data Source") +
  theme(title = element_text(face = 'bold.italic', color = "darkslategrey"),
        axis.title = element_text(face = 'plain', color = "darkslategrey")) +
  facet_wrap(~Source)
```



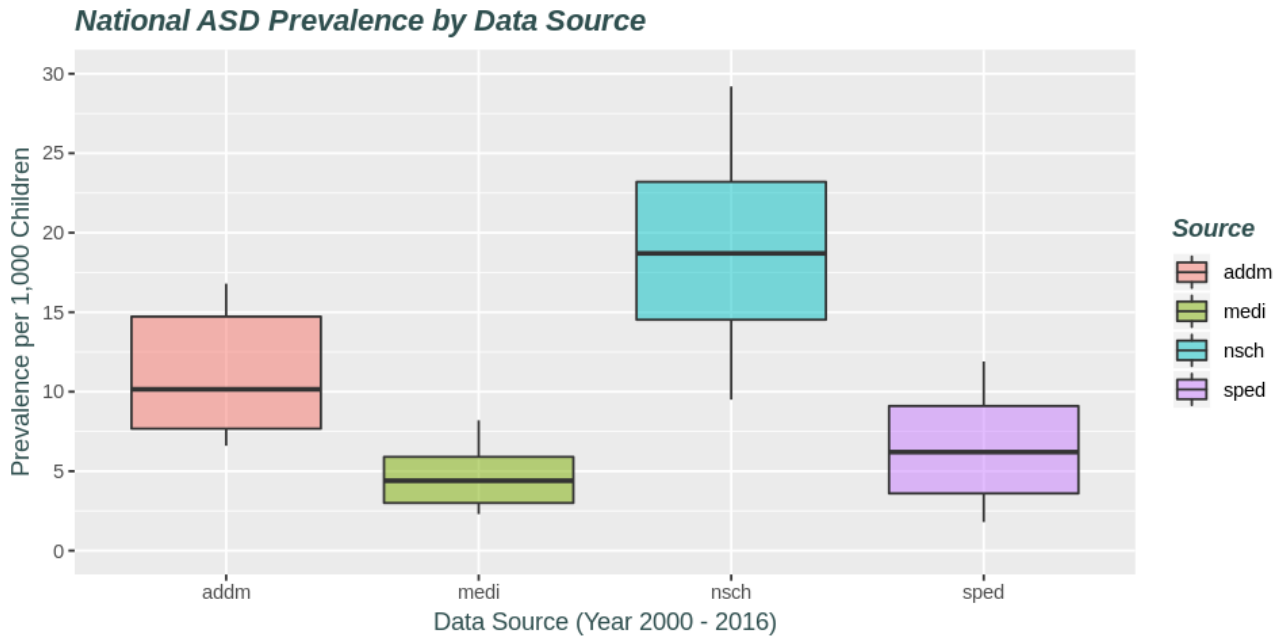
Data Visualisation (Enhanced) - Box plot

In [47]:
```
# Adjust in-line plot size to M x N
options(repr.plot.width=8, repr.plot.height=4)
```

In [48]:
```
# Create plot using R graphics
# 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")
```



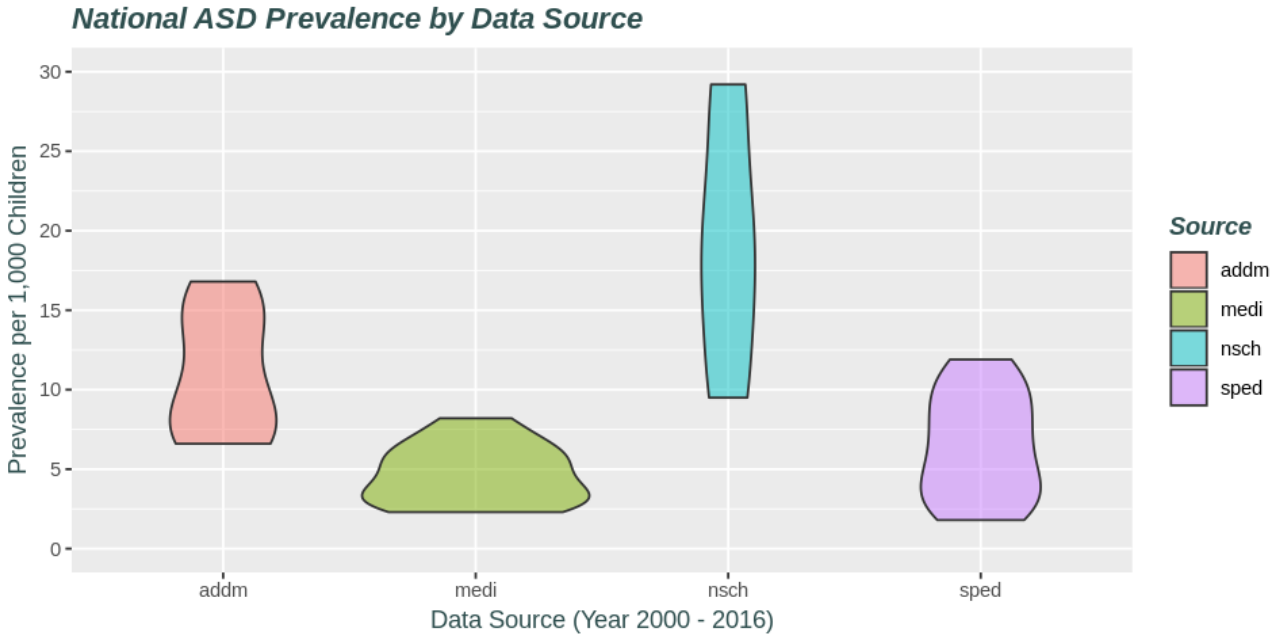
In [49]:
```
# Create box plot using ggplot2
ggplot(ASD_National, aes(x = Source, y = Prevalence, fill = Source)) +
  geom_boxplot(alpha = 0.5) +
  scale_y_continuous(name = "Prevalence per 1,000 Children",
                     breaks = seq(0, 30, 5),
                     limits=c(0, 30)) +
  scale_x_discrete(name = "Data Source (Year 2000 - 2016)") +
  ggtitle("National ASD Prevalence by Data Source") +
  theme(title = element_text(face = 'bold.italic', color = "darkslategrey"),
        axis.title = element_text(face = 'plain', color = "darkslategrey"))
```



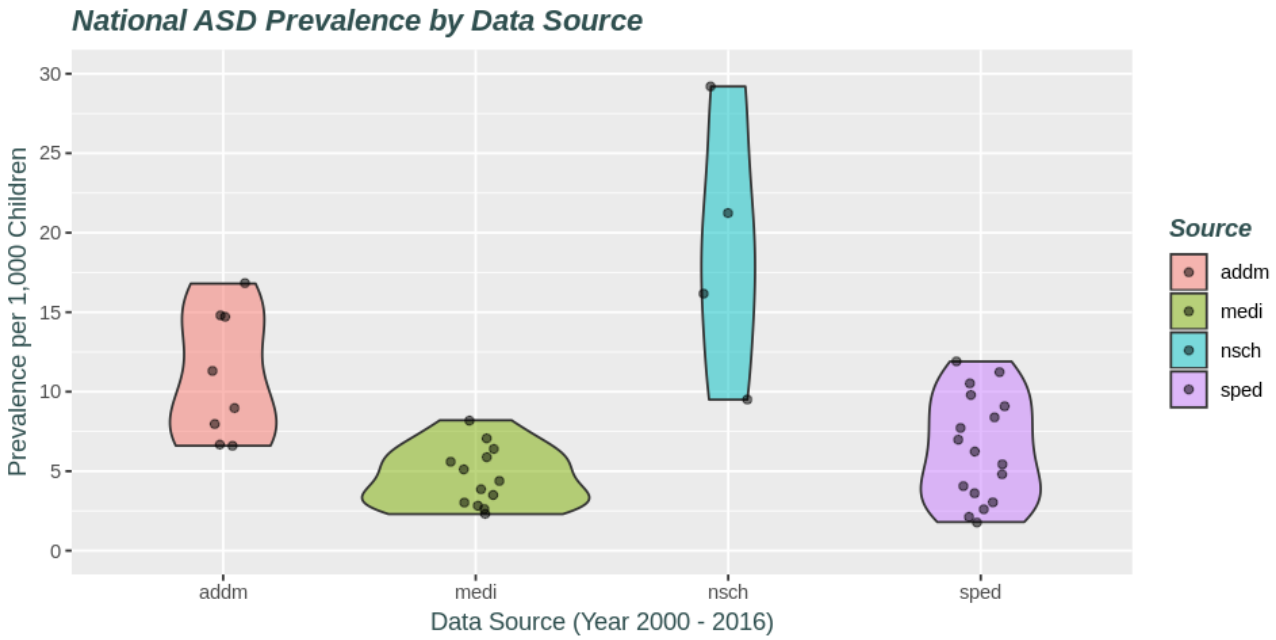
Data Visualisation (Enhanced) - Violin plot

In [50]:
```
# Adjust in-line plot size to M x N
# options(repr.plot.width=8, repr.plot.height=4)
```

In [51]:
```
# Create plot using ggplot2
ggplot(ASD_National, aes(x = Source, y = Prevalence, fill = Source)) +
  geom_violin(alpha = 0.5) +
  scale_y_continuous(name = "Prevalence per 1,000 Children",
                     breaks = seq(0, 30, 5),
                     limits=c(0, 30)) +
  scale_x_discrete(name = "Data Source (Year 2000 - 2016)") +
  ggtitle("National ASD Prevalence by Data Source") +
  theme(title = element_text(face = 'bold.italic', color = "darkslategrey"),
        axis.title = element_text(face = 'plain', color = "darkslategrey"))
```

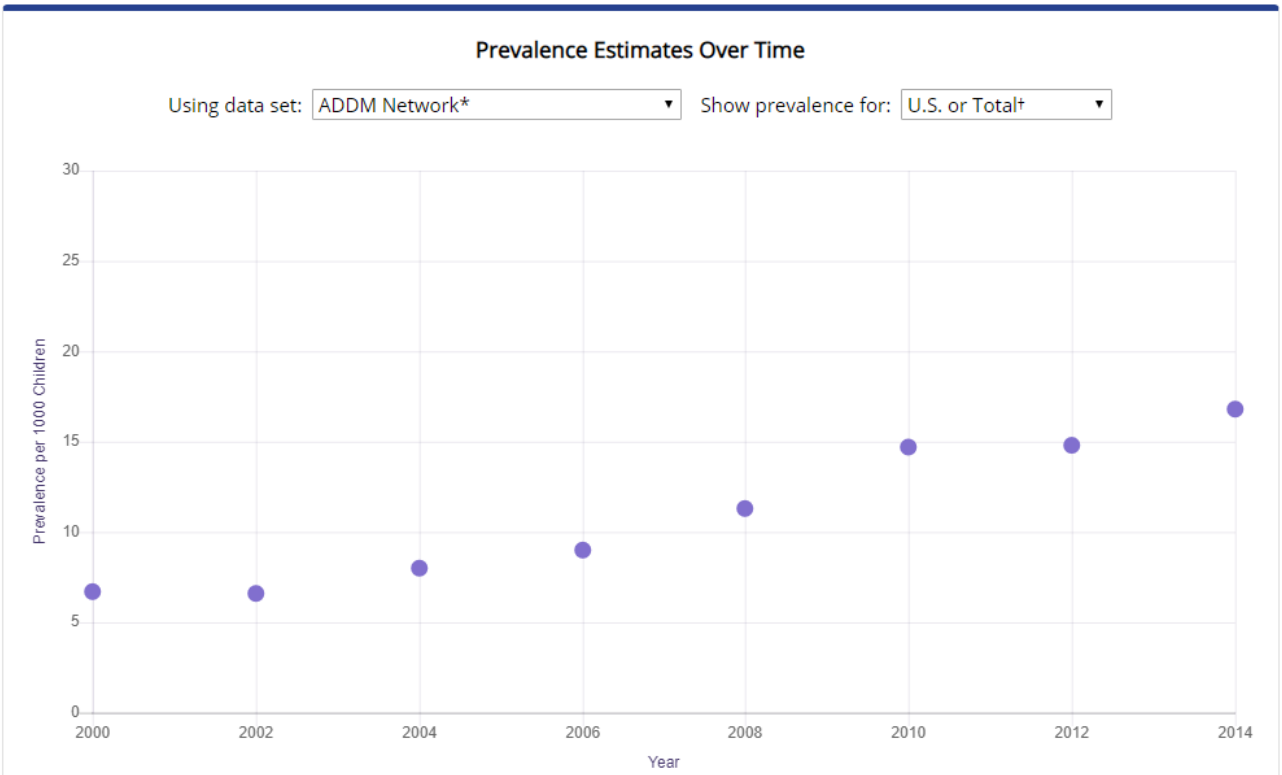


In [52]:
```
# Create plot using ggplot2
ggplot(ASD_National, aes(x = Source, y = Prevalence, fill = Source)) +
  geom_violin(alpha = 0.5) +
  geom_jitter(alpha = 0.5, position = position_jitter(width = 0.1)) + # Overla
#  coord_flip() + # Uncomment to flip x-y axis
  scale_y_continuous(name = "Prevalence per 1,000 Children",
                     breaks = seq(0, 30, 5),
                     limits=c(0, 30)) +
  scale_x_discrete(name = "Data Source (Year 2000 - 2016)") +
  ggtitle("National ASD Prevalence by Data Source") +
  theme(title = element_text(face = 'bold.italic', color = "darkslategrey"),
        axis.title = element_text(face = 'plain', color = "darkslategrey"))
```



Data Visualisation (Enhanced) - Line chart

Data Visualisation (Enhanced) - [CDC] REPORTED PREVALENCE HAS CHANGED OVER TIME

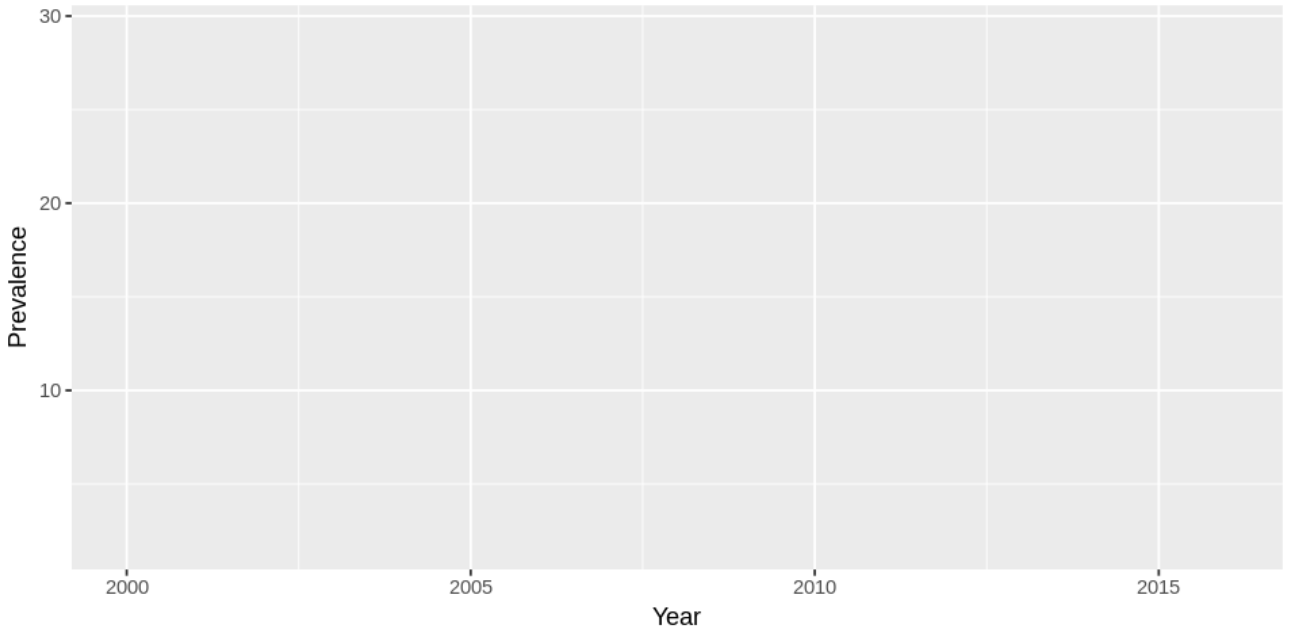


Note: *Hover your mouse over data points above to show prevalence by year.*
**ADDM Network data only represent a selection of sites within states (but not entire states) that were funded during each project cycle; therefore, data are not available for the entire United States.*
***ADDM estimate = the total for all sites combined.*
*†**NSCH data are not comparable over time as data collection methods changed. See technical notes for further details.***
††For NSCH data, data are suppressed when the width of the confidence interval exceeds 1.2 times the point estimate. This is the same approach that is recommended by NSCH here and our point estimates (and suppressed data points) match those on childhealthdata.org for 2016-2017.
†††If NSCH 2016 data are selected, combined 2016-2017 estimates are shown.

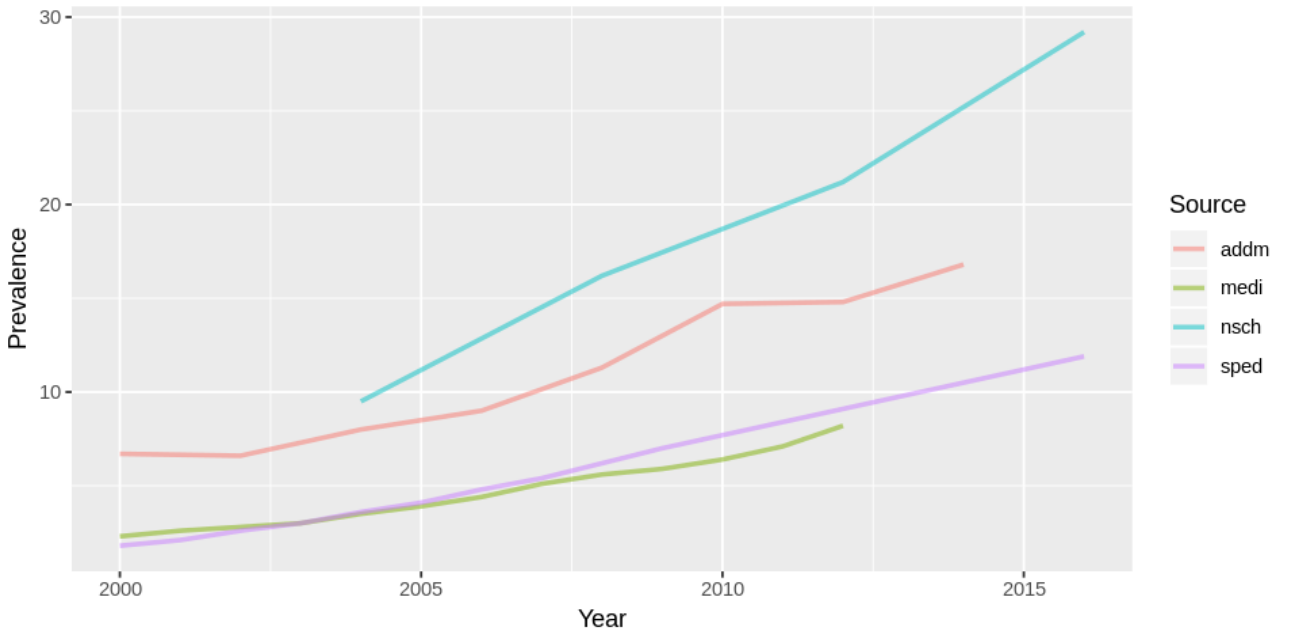
Data Visualisation (Enhanced) - [R] REPORTED PREVALENCE HAS CHANGED OVER TIME [Source: ALL]

In [53]:
```
# Adjust in-line plot size to M x N
options(repr.plot.width=8, repr.plot.height=4)
```

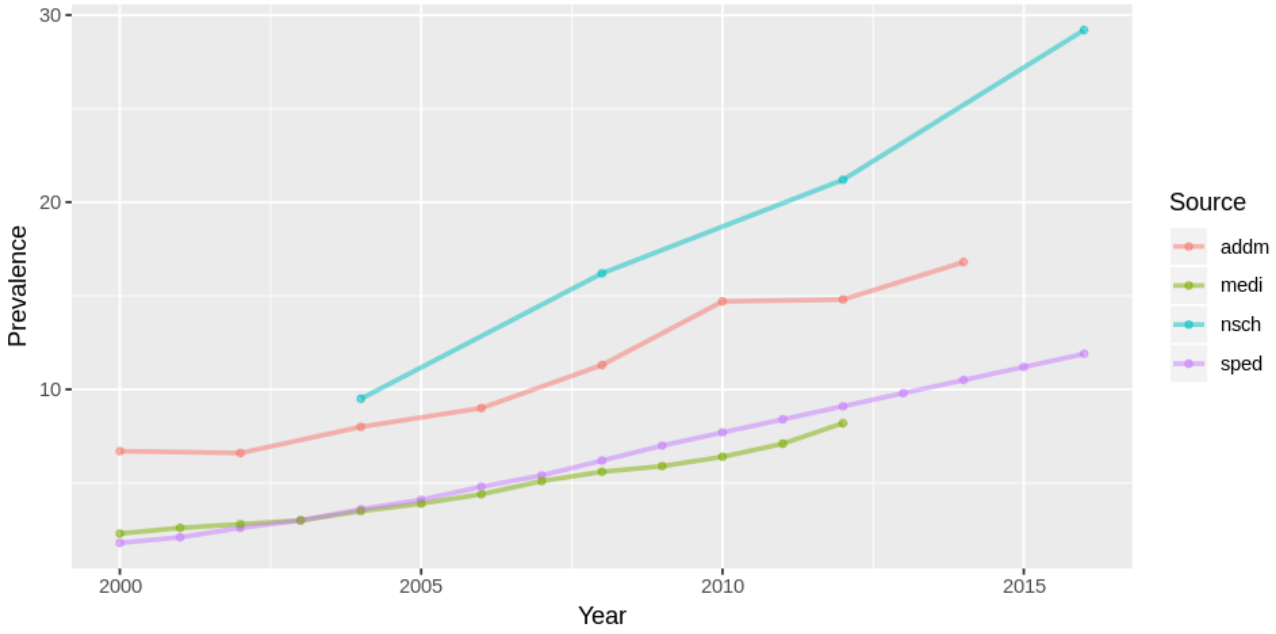
In [54]:
```
# ---------------------------------
# Build chart/plot layer by layer
# ---------------------------------

# Define a ggplot graphic object; provide data and x y for use
p <- ggplot(ASD_National, aes(x = Year, y = Prevalence))
# Show plot
p
```



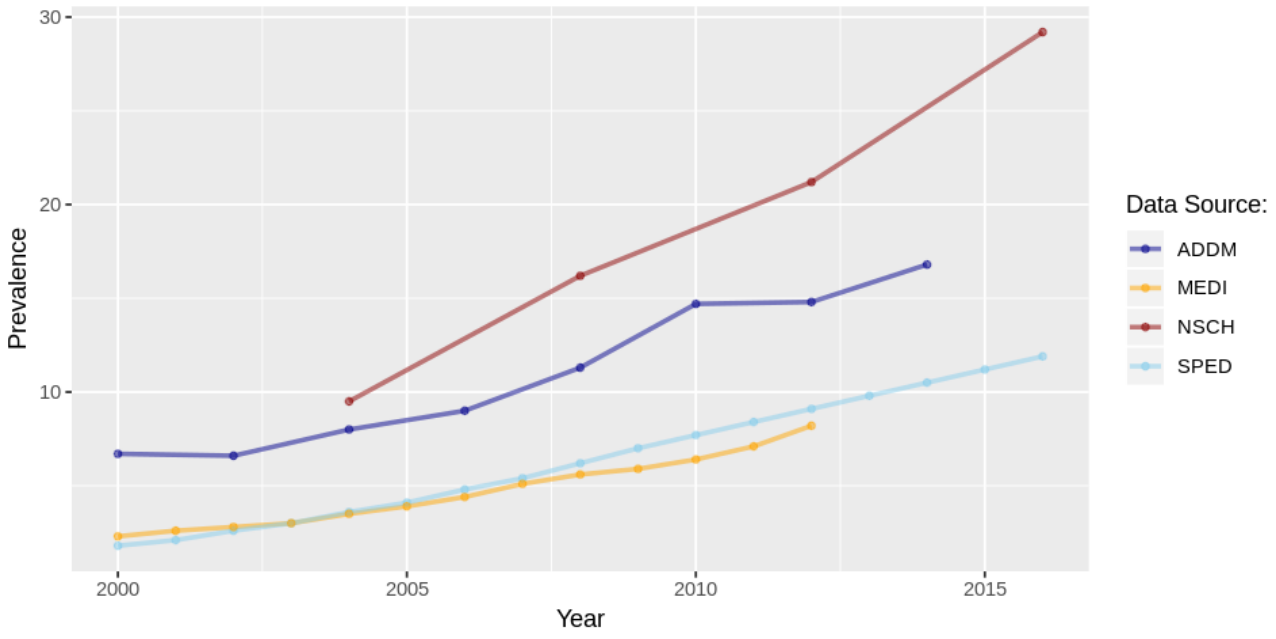
In [55]:
```
# Select (add) line chart type:
p <- p + geom_line(aes(color = Source),
                   linetype = "solid",  # http://sape.inf.usi.ch/quick-referen
                   size=1,
                   alpha=0.5)
# Show plot
p
```



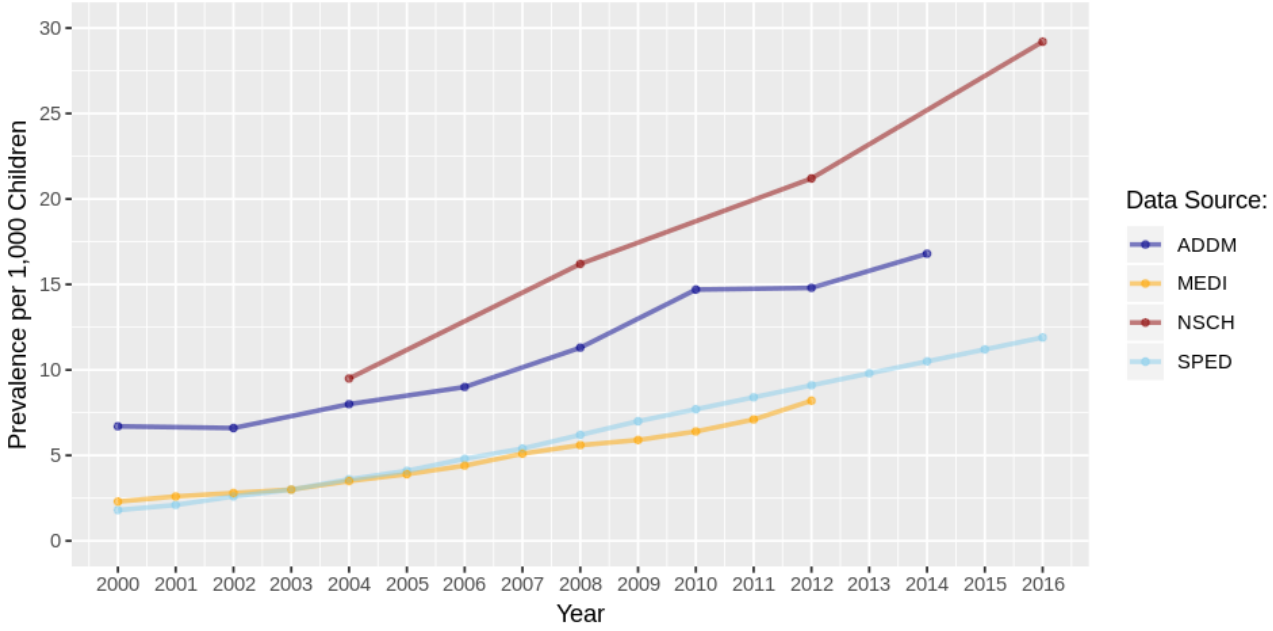
In [56]:
```
# Select (add) points to chart:
p <- p + geom_point(aes(color = Source),
                    size=2,
                    shape=20,
                    alpha=0.5)
# Show plot
p
```



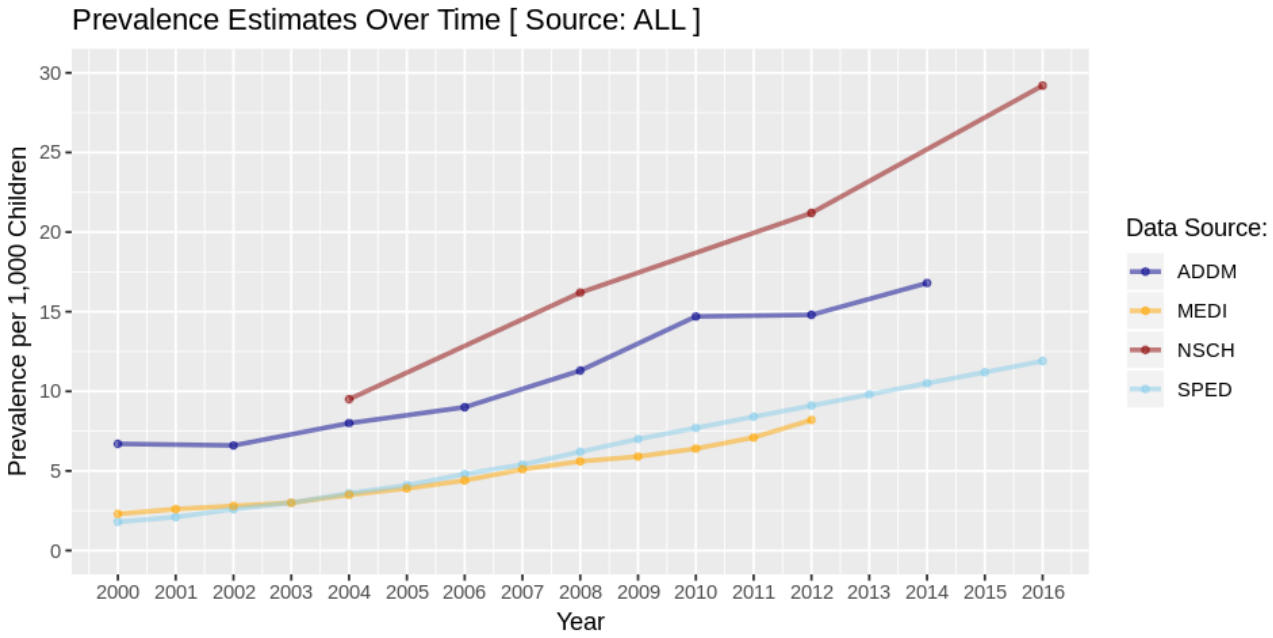
In [57]:
```
# Customize line color and legend name:
p <- p + scale_color_manual("Data Source:",
                            labels = c('ADDM', 'MEDI', 'NSCH', 'SPED'),
                            values = c("addm" = "darkblue",
                                       "medi" = "orange",
                                       "nsch" = "darkred",
                                       "sped" = "skyblue"))
# Show plot
p
```



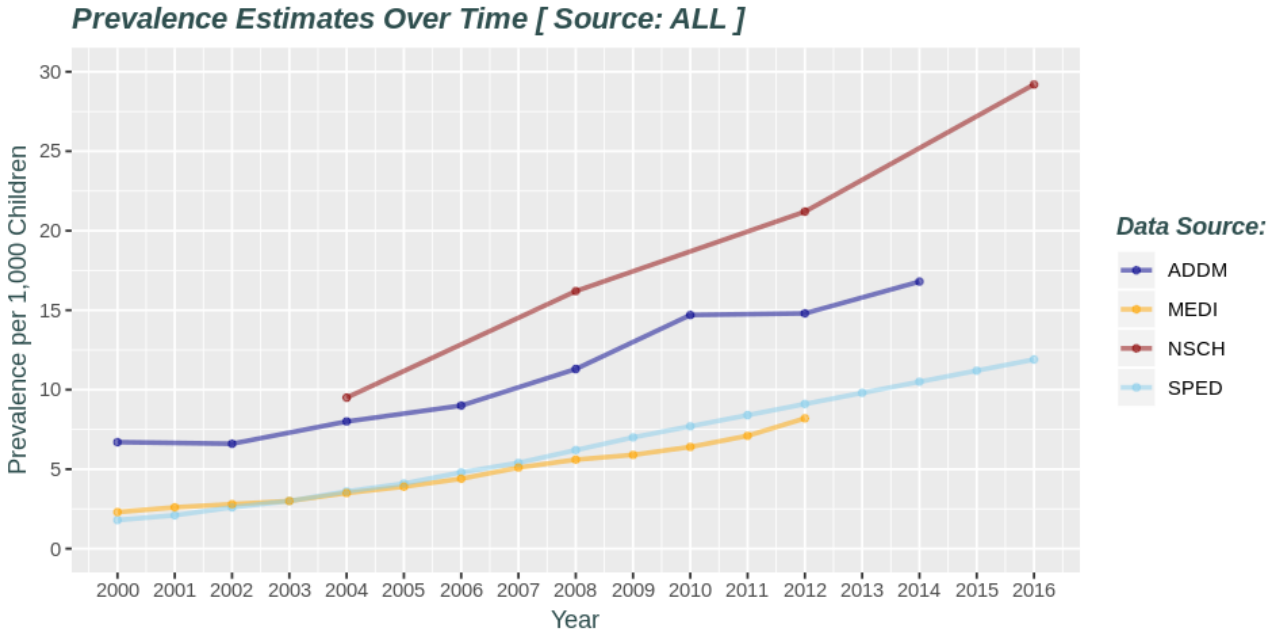
In [58]:
```
# Adjust x and y axis, scale, limit and labels:
p <- p + scale_y_continuous(name = "Prevalence per 1,000 Children",
                            breaks = seq(0, 30, 5),
                            limits=c(0, 30)) +
  scale_x_continuous(name = "Year",
                     breaks = seq(2000, 2016, 1),
                     limits = c(2000, 2016))
# Show plot
p
```



In [59]:
```
# Customise chart title:
p <- p + ggtitle("Prevalence Estimates Over Time [ Source: ALL ]")
# Show plot
p
```

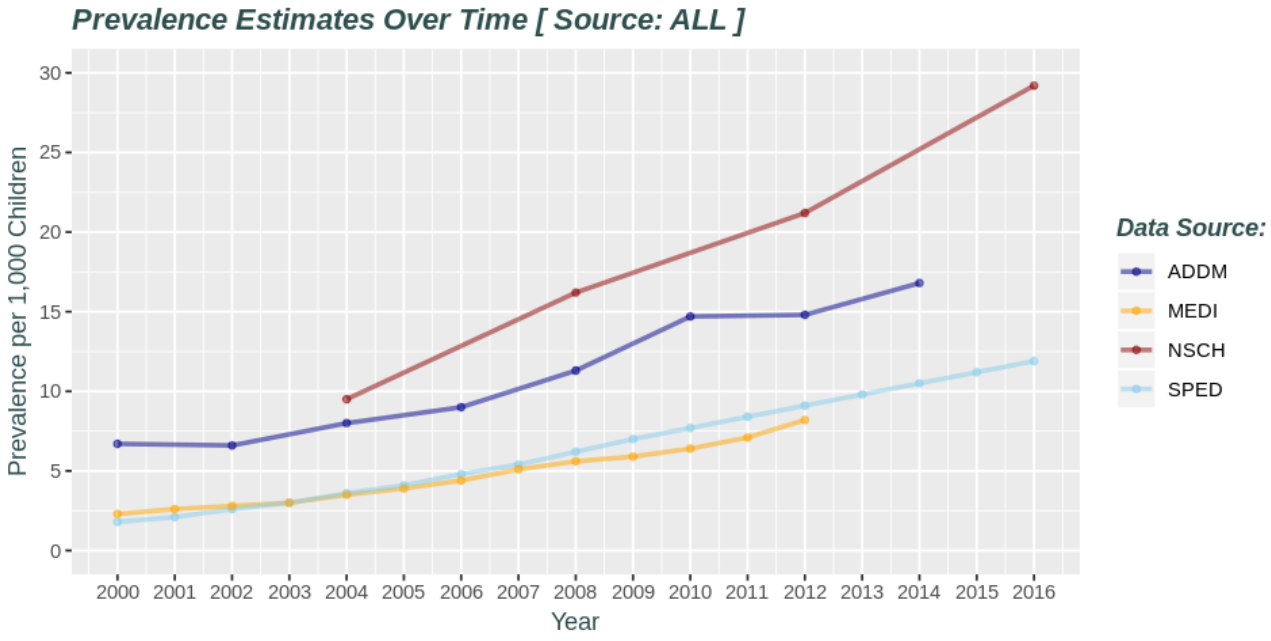


In [60]:
```
# Customise chart title and axis labels:
p <- p + theme(title = element_text(face = 'bold.italic', color = "darkslategr
               axis.title = element_text(face = 'plain', color = "darkslategre
# Show plot
p
```



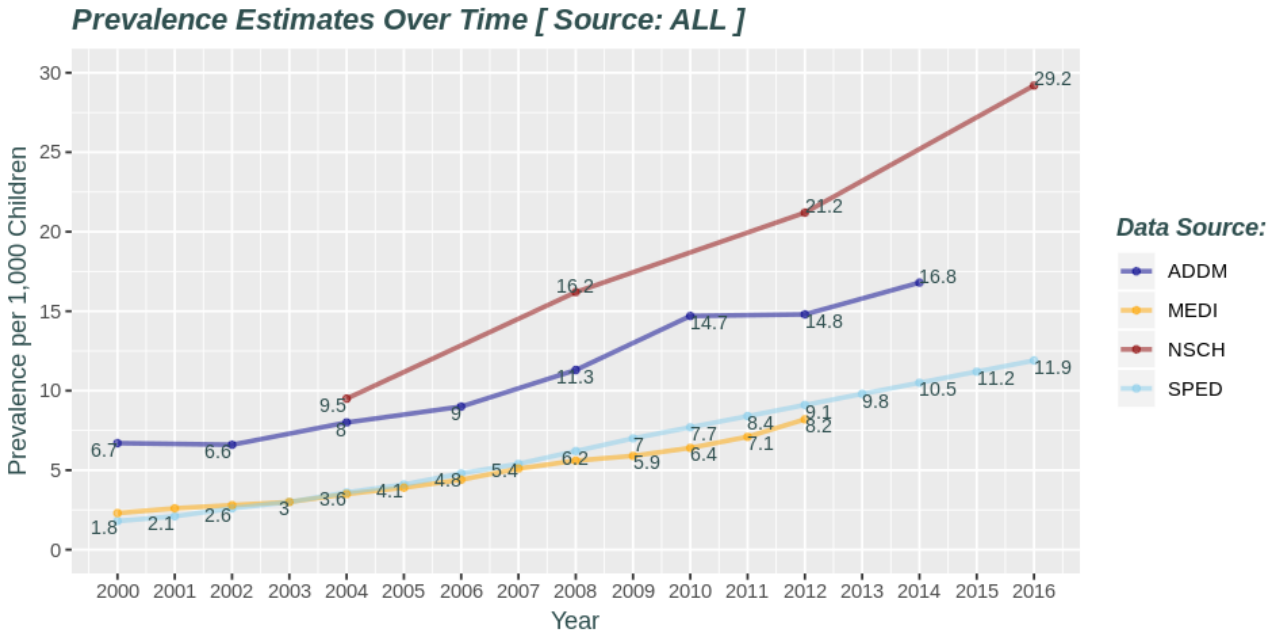
Consolidate above code into one chunk:

In [61]:
```
# ---------------------------------
# Consolidate above code into one chunk
# ---------------------------------
p <- ggplot(ASD_National, aes(x = Year, y = Prevalence)) +
  geom_line(aes(color = Source),
            linetype = "solid",  # http://sape.inf.usi.ch/quick-reference/ggpl
            size=1,
            alpha=0.5) +
  geom_point(aes(color = Source),
             size=2,
             shape=20,
             alpha=0.5) +
  scale_color_manual("Data Source:",
                     labels = c('ADDM', 'MEDI', 'NSCH', 'SPED'),
                     values = c("addm" = "darkblue",
                                "medi" = "orange",
                                "nsch" = "darkred",
                                "sped" = "skyblue")) +
  scale_y_continuous(name = "Prevalence per 1,000 Children",
                     breaks = seq(0, 30, 5),
                     limits=c(0, 30)) +
  scale_x_continuous(name = "Year",
                     breaks = seq(2000, 2016, 1),
                     limits = c(2000, 2016)) +
  ggtitle("Prevalence Estimates Over Time [ Source: ALL ]") +
  theme(title = element_text(face = 'bold.italic', color = "darkslategrey"),
        axis.title = element_text(face = 'plain', color = "darkslategrey"))
# Show plot
p
```



Optionally, display data values/labels:

In [62]:
```
# Optionally, displaydata values/labels
p + geom_text(aes(label = round(Prevalence, 1)), # Values are rounded for disp
              vjust = "outward",
              #          nudge_y = 0.2, # optionally life the text
              hjust = "outward",
              check_overlap = TRUE,
              size = 3, # size of textual data label
              col = 'darkslategrey')
```



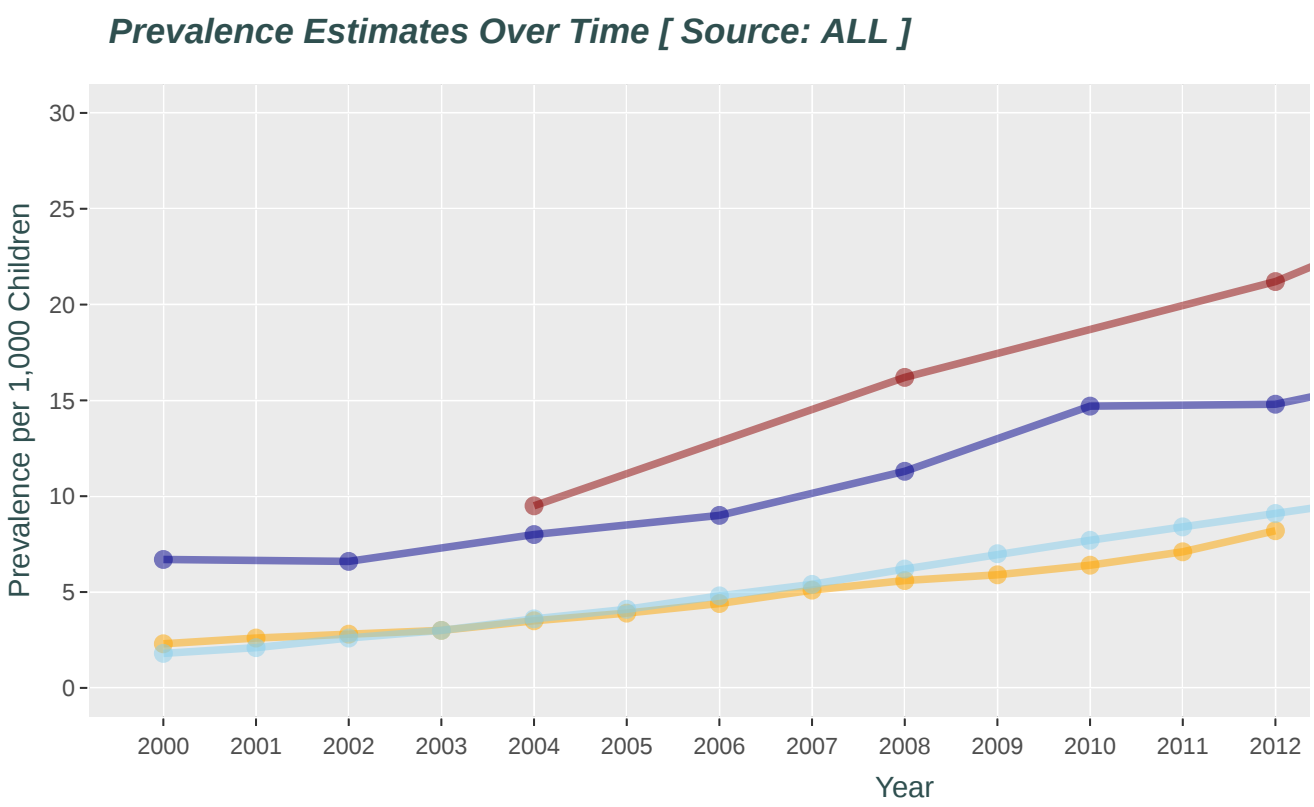
Data Visualisation (Enhanced) - Dynamic Visualisation with plotly

In [63]:
```
if(!require(plotly)){install.packages("plotly")}
library(plotly)
```

```
Loading required package: plotly

Attaching package: ‘plotly’

The following object is masked from ‘package:ggplot2’:

    last_plot

The following object is masked from ‘package:stats’:

    filter

The following object is masked from ‘package:graphics’:

    layout
```

Create ploly graph object from ggplot graph object:

In [64]:
```
p_dynamic <- p
p_dynamic <- ggplotly(p_dynamic)
p_dynamic
```



Data Visualisation (Enhanced) - Use themes as aesthetic template

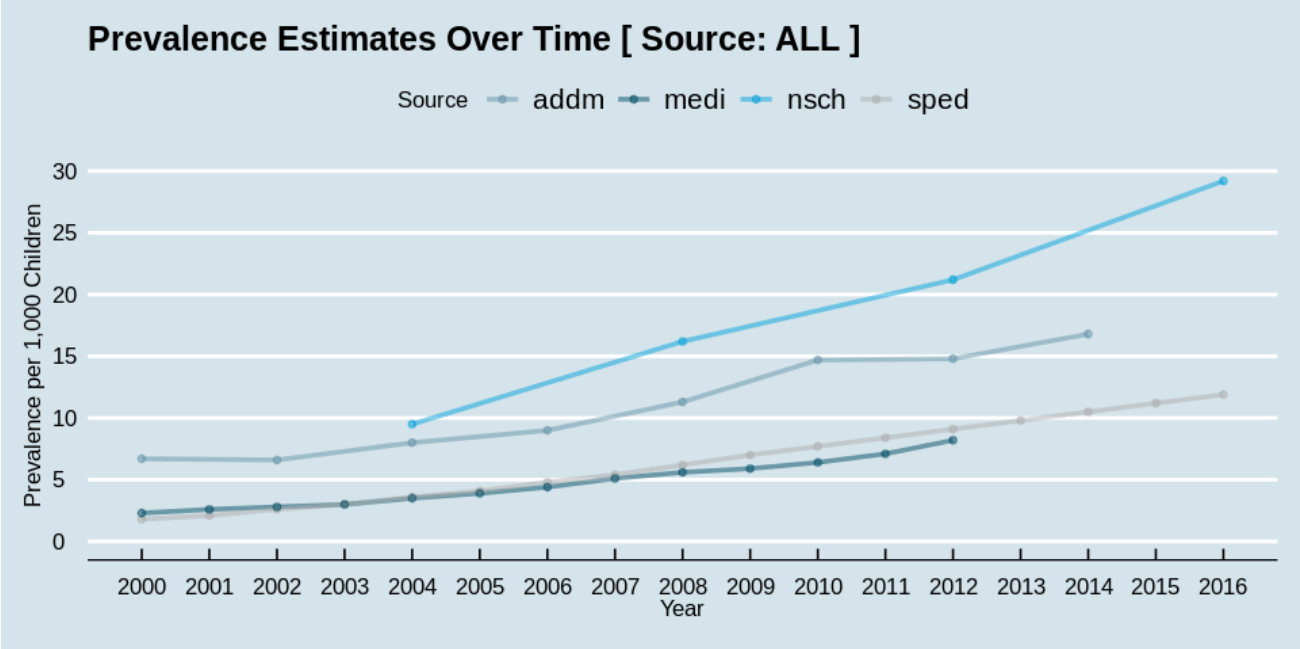
In [65]:
```
if(!require(ggthemes)){install.packages("ggthemes")}
library('ggthemes')
```

```
Loading required package: ggthemes
```

Theme of the Economist magazine:

In [66]:
```
# Theme of the economist magazine:
p + theme_economist() + scale_colour_economist()
```

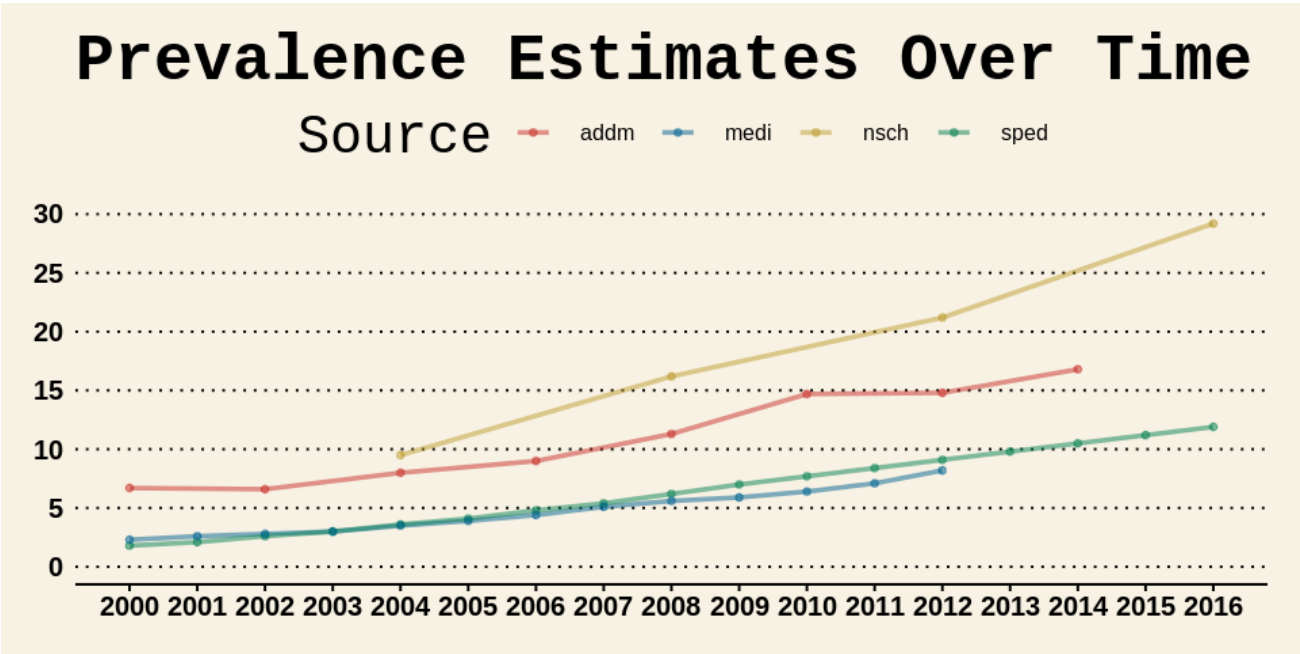
```
Scale for 'colour' is already present. Adding another scale for 'colour',
which will replace the existing scale.
```



Theme of the Wall Street Journal:

In [67]:
```
# Theme of the Wall Street Journal:
p + theme_wsj() + scale_colour_wsj("colors6")
```

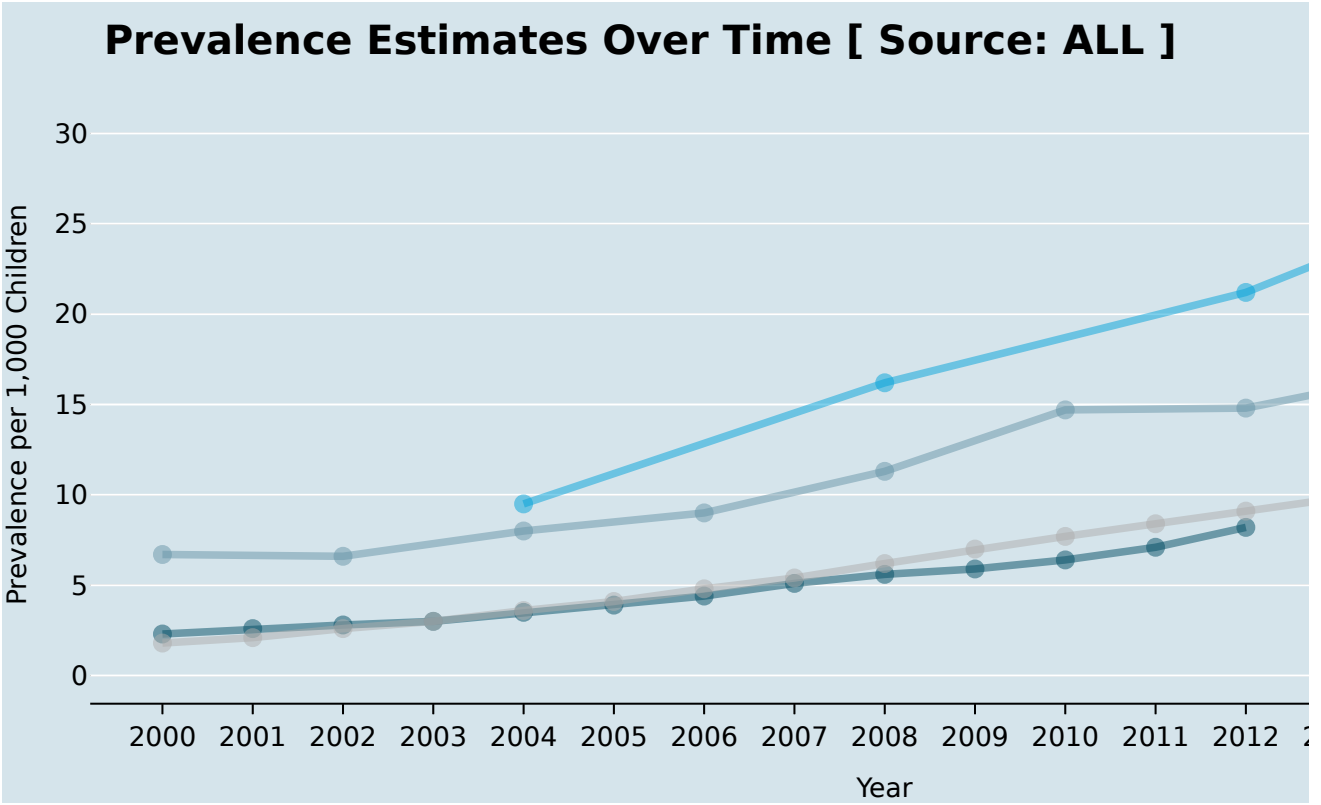
```
Scale for 'colour' is already present. Adding another scale for 'colour',
which will replace the existing scale.
```



Dynamic chart with theme of the economist magazine:

In [68]:
```
# Dynamic chart with theme of the economist magazine:
p_dynamic <- p + theme_economist() + scale_colour_economist()
p_dynamic <- ggplotly(p_dynamic)
p_dynamic
```

```
Scale for 'colour' is already present. Adding another scale for 'colour',
which will replace the existing scale.
```



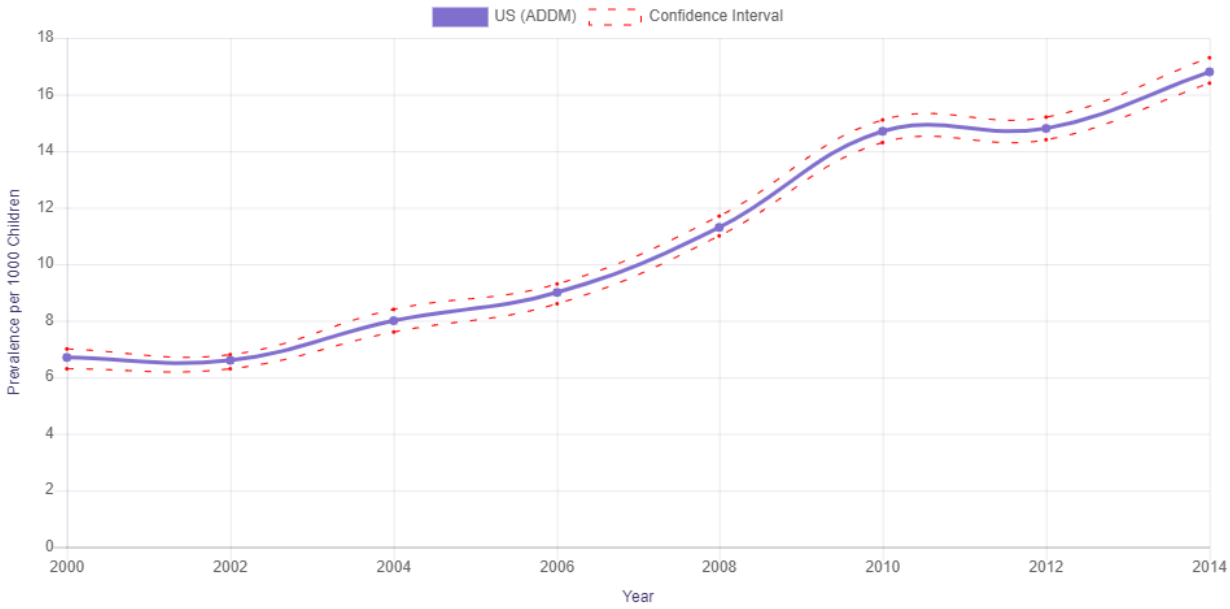
Data Visualisation (Enhanced) - [CDC] ADDM Network estimates for overall ASD prevalence in US over time [Source: ADDM] over [Year]

ADDM NETWORK DATA

In this section, explore the most recent ADDM data, both overall and among certain demographic groups by study area.

ADDM Network estimates for overall ASD prevalence in US over time

with confidence interval



**ADDM data do not represent the entire state, only a selection of sites within the state.*

***ADDM estimate = the total for all sites combined.*

†NSCH data are not comparable over time as data collection methods changed and the data are not provided here. See technical notes for further details.

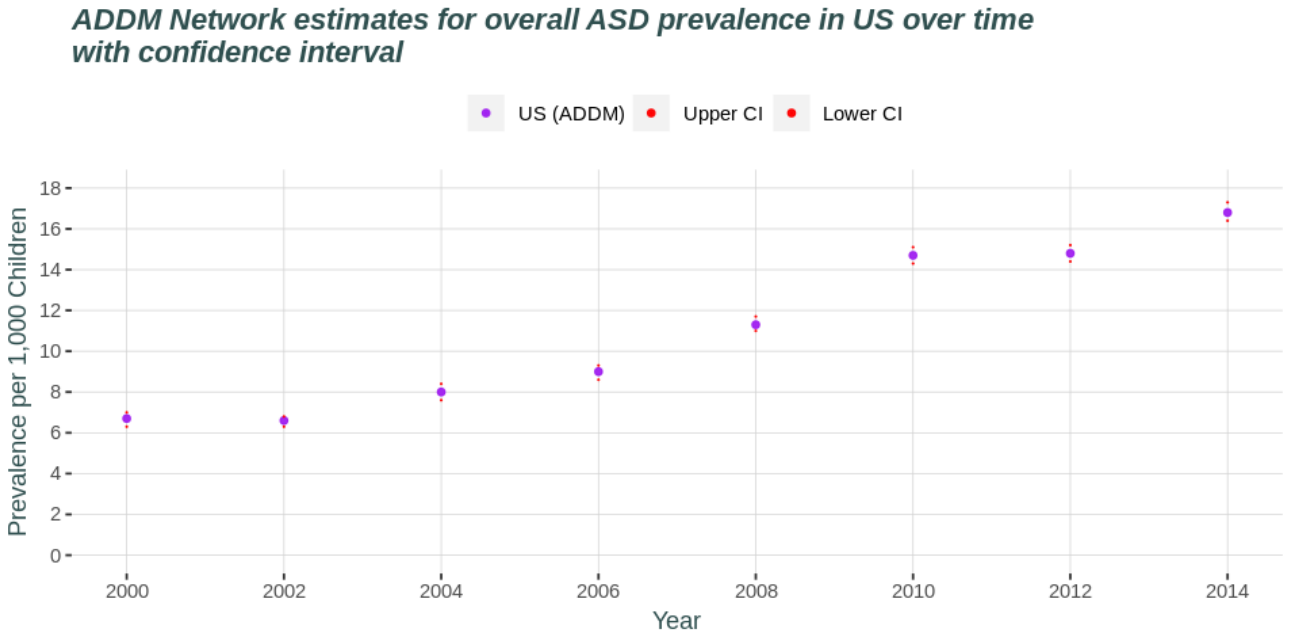
Data Visualisation (Enhanced) - [R] ADDM Network estimates for overall ASD prevalence in US over time [Source: ADDM] over [Year]

In [69]:
```r
# Adjust in-line plot size to M x N
options(repr.plot.width=8, repr.plot.height=4)
```

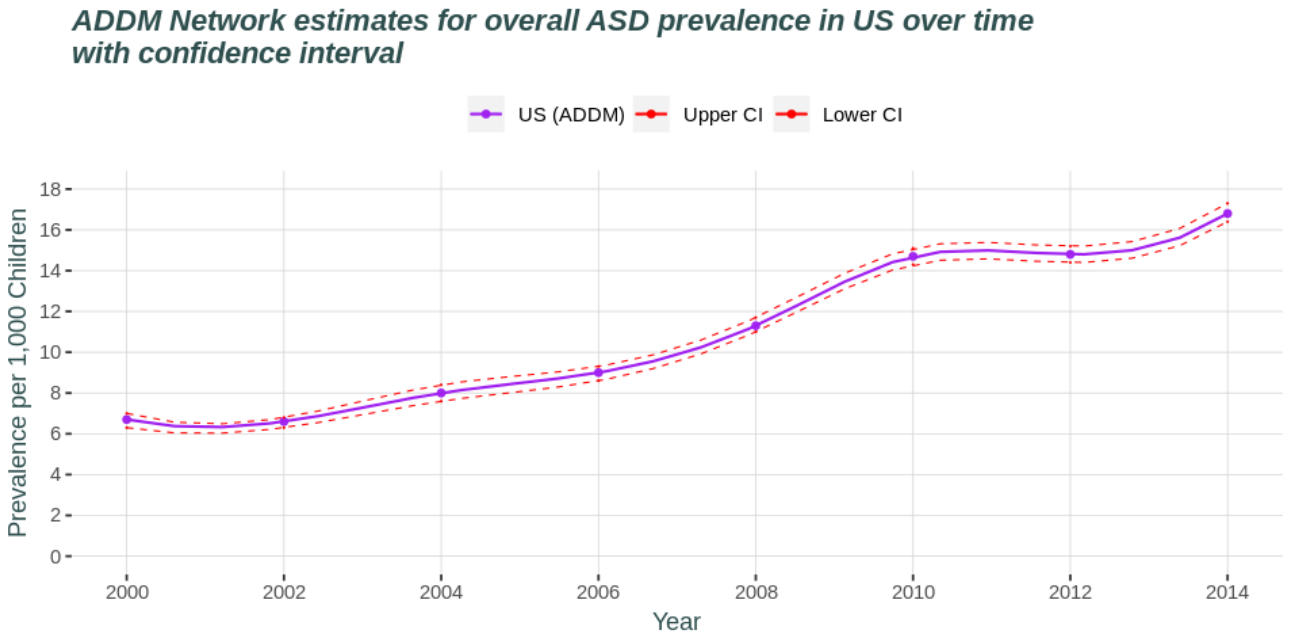
In [70]:
```r
# Filter only data of ADDM
ASD_National_ADDM <- subset(ASD_National, Source == 'addm')
```

In [71]:

```
# ---------------------------------
# [addm] ADDM Network estimates for overall ASD prevalence in US over time
# ---------------------------------

#  Color:
# 'ADDM_Average' "purple"

p <- ggplot(ASD_National_ADDM, aes(x = Year, y = Prevalence)) +
  geom_point(aes(y = Prevalence, color = 'ADDM_Average'), # Name for manual co
             size=2,
             shape=20,
             alpha=0.95) +
  # Add point for Upper.CI
  geom_point(aes(y = Upper.CI, color = 'ADDM_U_CI'), # Name for manual colour
             size=0.1,
             shape=20,
             alpha=0.95) +
  # Add point for Upper.CI
  geom_point(aes(y = Lower.CI, color = 'ADDM_L_CI'), # Name for manual colour
             size=0.1,
             shape=20,
             alpha=0.95) +
  scale_colour_manual(name="",
                      labels = c("US (ADDM)", "Upper CI", "Lower CI"), # Names
                      values = c(ADDM_Average="purple", ADDM_U_CI="red", ADDM_
# Add title, axis label, and axis scale
p <- p + scale_y_continuous(name = "Prevalence per 1,000 Children",
                            breaks = seq(0, 18, 2),
                            limits=c(0, 18)) +
  scale_x_continuous(name = "Year",
                     breaks = seq(2000, 2014, 2),
                     limits = c(2000, 2014)) +
  ggtitle("ADDM Network estimates for overall ASD prevalence in US over time\n
  theme(title = element_text(face = 'bold.italic', color = "darkslategrey"),
        axis.title = element_text(face = 'plain', color = "darkslategrey"),
        panel.background = element_blank(), # Remove chart backgroun colour
        legend.position = 'top',
        panel.grid.major = element_line(size = 0.2, linetype = 'solid', colour
       )
# Show plot
p
```



In [72]:
```
# Add smooth curve to go through date points, using interpolation with splines
# https://stackoverflow.com/questions/35205795/plotting-smooth-line-through-al
spline_ADDM_Prevalence <- as.data.frame(spline(ASD_National_ADDM$Year, ASD_Nat
spline_ADDM_Prevalence_U_CI <- as.data.frame(spline(ASD_National_ADDM$Year, AS
spline_ADDM_Prevalence_L_CI <- as.data.frame(spline(ASD_National_ADDM$Year, AS
# Show plot
p + geom_line(data = spline_ADDM_Prevalence, aes(x = x, y = y, color = 'ADDM_A
  geom_line(data = spline_ADDM_Prevalence_U_CI, aes(x = x, y = y, color = 'ADD
  geom_line(data = spline_ADDM_Prevalence_L_CI, aes(x = x, y = y, color = 'ADD
```



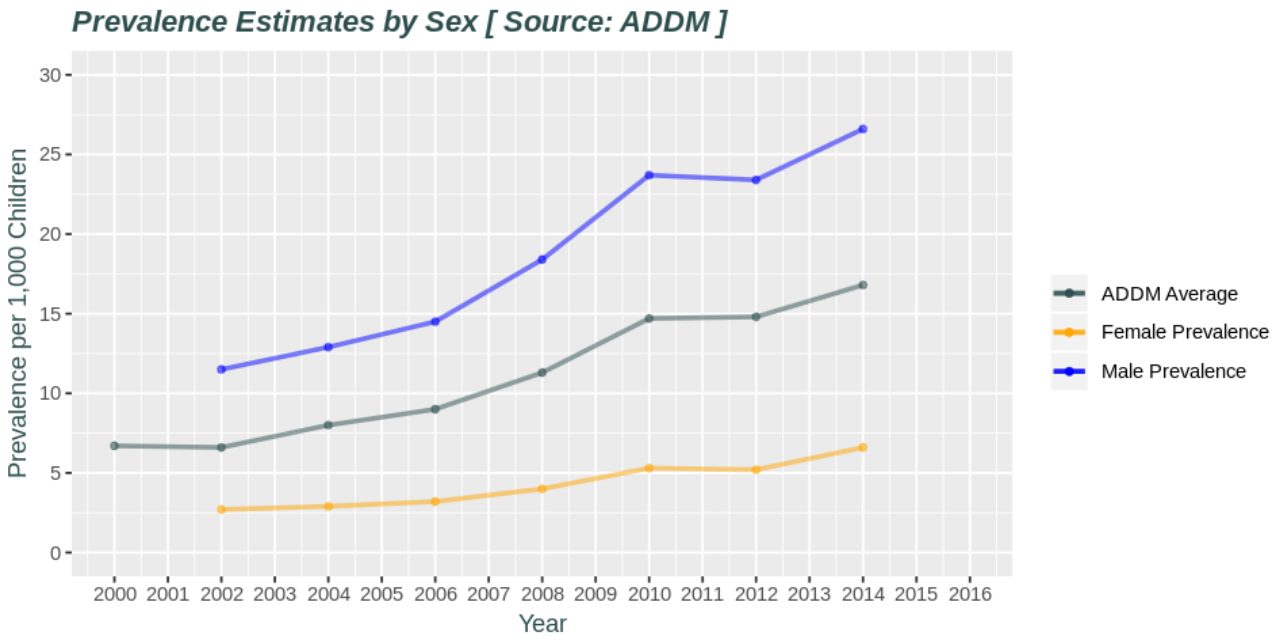
Data Visualisation (Enhanced) - [R] REPORTED PREVALENCE VARIES BY SEX [Source: ADDM] over [Year]

In [73]:
```
# Adjust in-line plot size to M x N
# options(repr.plot.width=8, repr.plot.height=4)
```

In [74]:

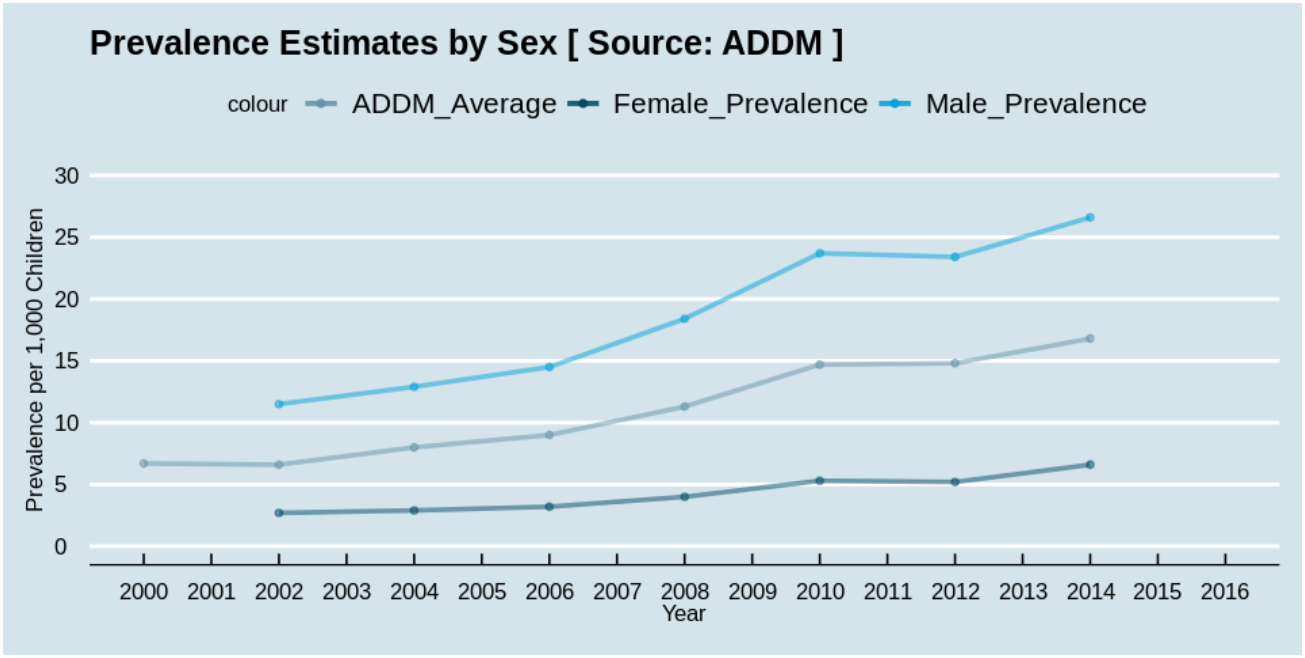
```
# ---------------------------------
# [addm] < Prevalence Varies by Sex >
# ---------------------------------

#  Color:
# 'ADDM_Average' "darkslategrey"
# 'Female_Prevalence' "orange"
# 'Male_Prevalence' "blue"

p <- ggplot(ASD_National_ADDM, aes(x = Year, y = Prevalence)) +
  geom_line(aes(y = Prevalence, colour = 'ADDM_Average'),
            linetype = "solid",  # http://sape.inf.usi.ch/quick-reference/ggpl
            size=1,
            alpha=0.5) +
  geom_point(aes(y = Prevalence, color = 'ADDM_Average'),
             size=2,
             shape=20,
             alpha=0.5) +
  # Add line for Female
  geom_line(aes(y = Female.Prevalence, colour = 'Female_Prevalence'),
            linetype = "solid",  # http://sape.inf.usi.ch/quick-reference/ggpl
            size=1,
            alpha=0.5) +
  geom_point(aes(y = Female.Prevalence, color = 'Female_Prevalence'),
             size=2,
             shape=20,
             alpha=0.5) +
  # Add line for Male
  geom_line(aes(y = Male.Prevalence, colour = 'Male_Prevalence'),
            linetype = "solid",  # http://sape.inf.usi.ch/quick-reference/ggpl
            size=1,
            alpha=0.5) +
  geom_point(aes(y = Male.Prevalence, color = 'Male_Prevalence'),
             size=2,
             shape=20,
             alpha=0.5) +
  scale_colour_manual(name="",
                      labels = c("ADDM Average", "Female Prevalence", "Male Pr
                      values = c(ADDM_Average="darkslategrey", Female_Prevalen
# Add title, axis label, and axis scale
p <- p + scale_y_continuous(name = "Prevalence per 1,000 Children",
                            breaks = seq(0, 30, 5),
                            limits=c(0, 30)) +
  scale_x_continuous(name = "Year",
                     breaks = seq(2000, 2016, 1),
                     limits = c(2000, 2016)) +
  ggtitle("Prevalence Estimates by Sex [ Source: ADDM ]") +
  theme(title = element_text(face = 'bold.italic', color = "darkslategrey"),
        axis.title = element_text(face = 'plain', color = "darkslategrey"))
# Show plot
p
```

```
Warning message:
“Removed 1 rows containing missing values (geom_path).”Warning message:
“Removed 1 rows containing missing values (geom_point).”Warning message:
“Removed 1 rows containing missing values (geom_path).”Warning message:
“Removed 1 rows containing missing values (geom_point).”
```



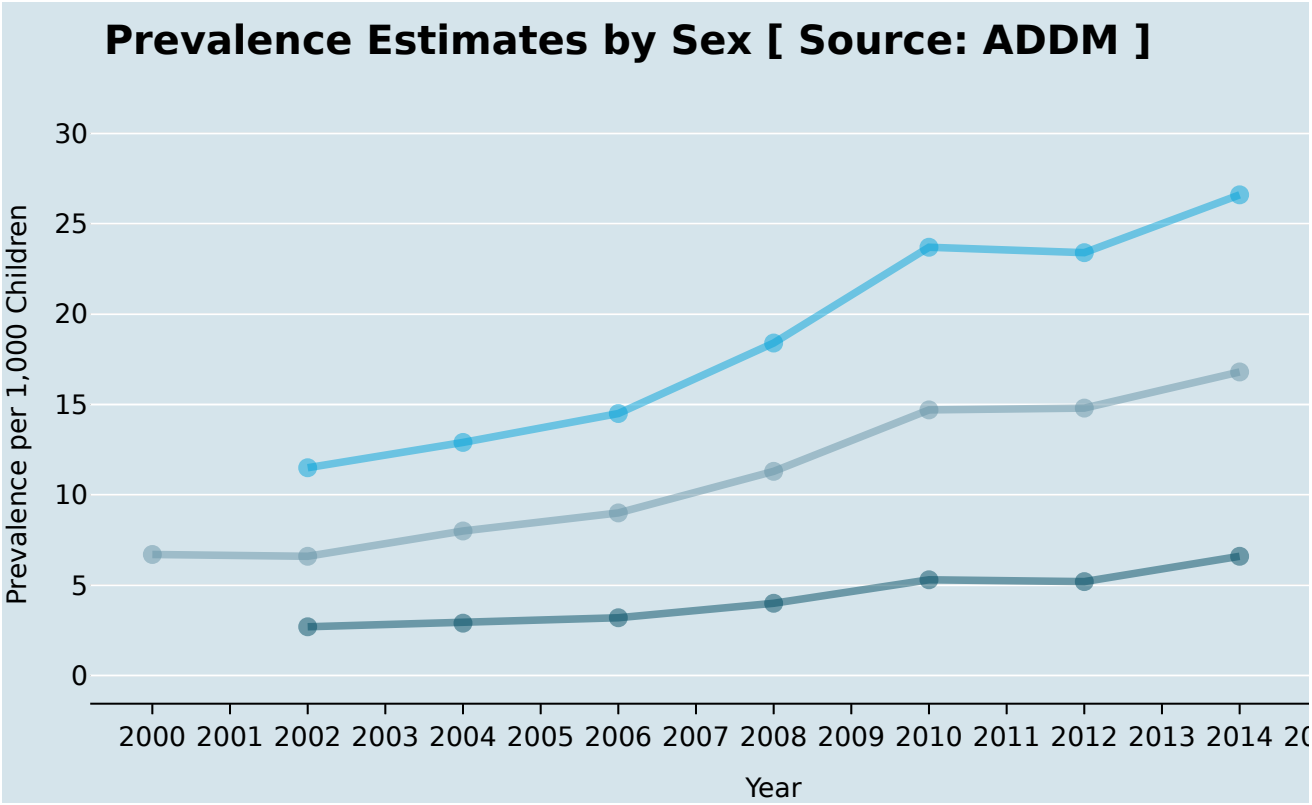
In [75]:
```
# Apply theme
p + theme_economist() + scale_colour_economist() # p + theme_wsj() + scale_col
```

```
Scale for 'colour' is already present. Adding another scale for 'colour',
which will replace the existing scale.
Warning message:
“Removed 1 rows containing missing values (geom_path).”Warning message:
“Removed 1 rows containing missing values (geom_point).”Warning message:
“Removed 1 rows containing missing values (geom_path).”Warning message:
“Removed 1 rows containing missing values (geom_point).”
```



In [76]:
```
# Dynamic chart:
p_dynamic <- p + theme_economist() + scale_colour_economist()
p_dynamic <- ggplotly(p_dynamic)
p_dynamic
```

```
Scale for 'colour' is already present. Adding another scale for 'colour',
which will replace the existing scale.
```



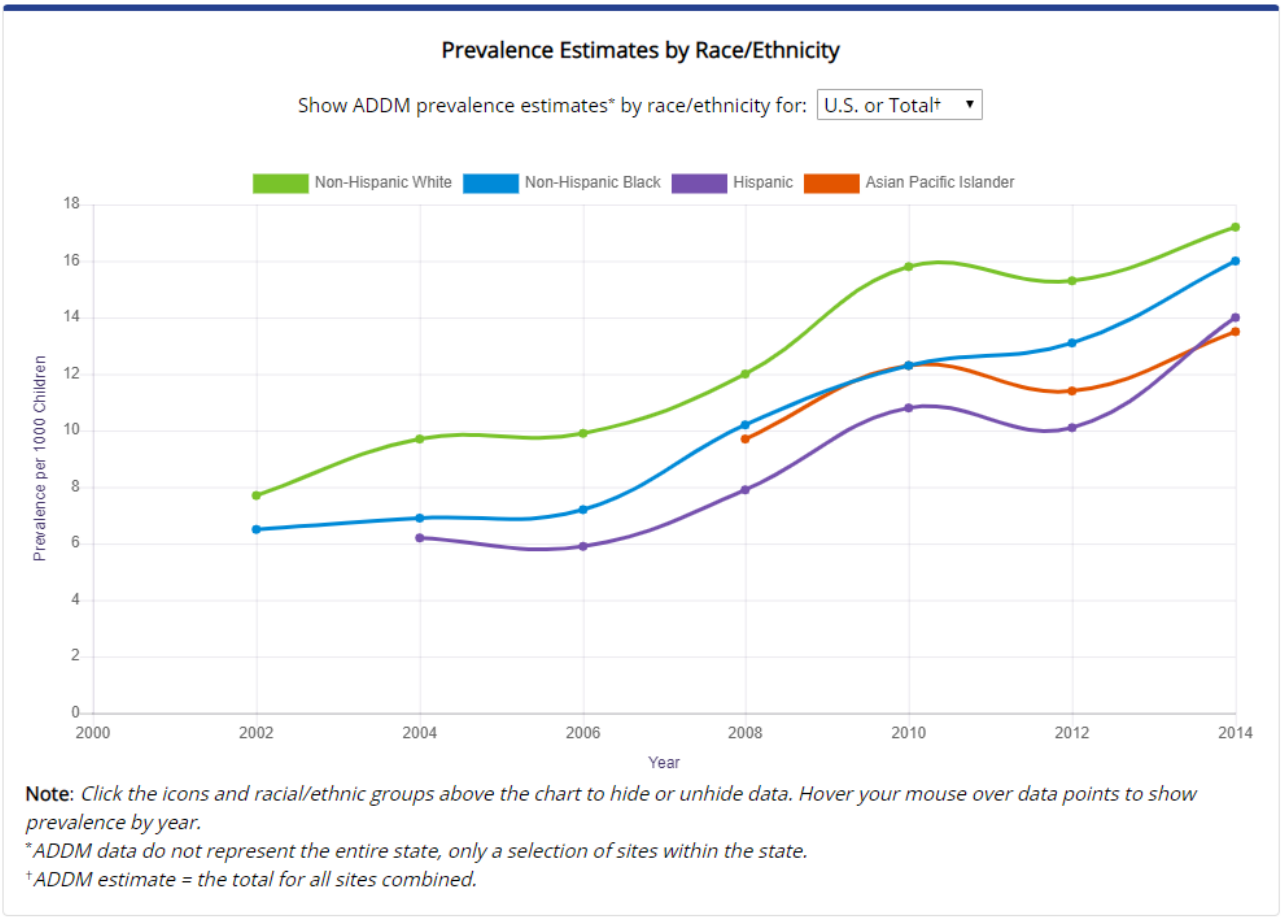
Quiz:

Add 95% Confidence Interval to above plot (Use ggplot)

In [77]:
```
# Write your code below and press Shift+Enter to execute
```

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

Data Visualisation (Enhanced) - [CDC] REPORTED PREVALENCE VARIES BY RACE AND ETHNICITY



Note: *Click the icons and racial/ethnic groups above the chart to hide or unhide data. Hover your mouse over data points to show prevalence by year.*

**ADDM data do not represent the entire state, only a selection of sites within the state.*

†ADDM estimate = the total for all sites combined.

Data Visualisation (Enhanced) - [R] REPORTED PREVALENCE VARIES BY RACE AND ETHNICITY [Source: ADDM] With Average

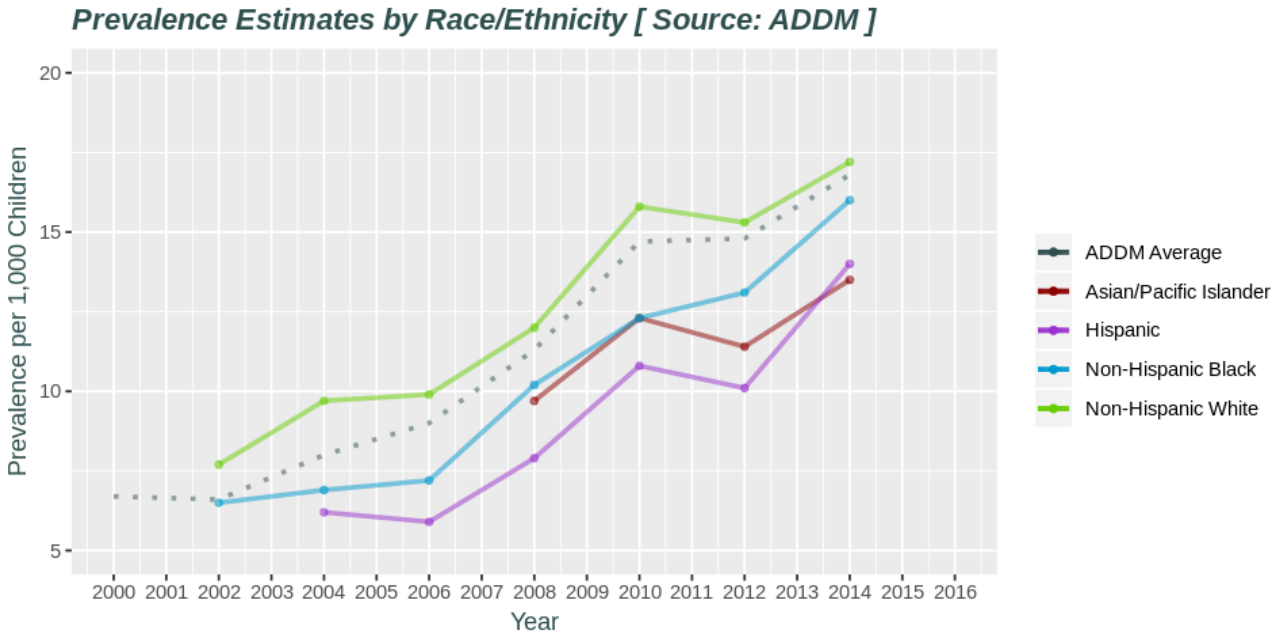
In [78]:
```
# Adjust in-line plot size to M x N
# options(repr.plot.width=8, repr.plot.height=4)
```

In [79]:

```
# ---------------------------------
# [addm] < Prevalence Varies by Race and Ethnicity >
# ---------------------------------

#  Color:
# 'ADDM_Average' "darkslategrey"
# 'Asian_Pacific_Islander' "darkred"
# 'Hispanic' "darkorchid3"
# 'Non_Hispanic_Black' "deepskyblue3"
# 'Non_Hispanic_White' "chartreuse3"

p <- ggplot(ASD_National_ADDM, aes(x = Year, y = Prevalence)) +
  geom_line(aes(y = Prevalence, colour = 'ADDM_Average'),
            linetype = "dotted",  # http://sape.inf.usi.ch/quick-reference/ggp
            size=1,
            alpha=0.5) +
  geom_point(aes(y = Prevalence, color = 'ADDM_Average'),
             size=2,
             shape=20,
             alpha=0) +
  # Add line for Asian.or.Pacific.Islander.Prevalence
  geom_line(aes(y = Asian.or.Pacific.Islander.Prevalence, colour = 'Asian_Paci
            linetype = "solid",  # http://sape.inf.usi.ch/quick-reference/ggpl
            size=1,
            alpha=0.5) +
  geom_point(aes(y = Asian.or.Pacific.Islander.Prevalence, colour = 'Asian_Pac
             size=2,
             shape=20,
             alpha=0.5) +
  # Add line for Hispanic.Prevalence
  geom_line(aes(y = Hispanic.Prevalence, colour = 'Hispanic'),
            linetype = "solid",  # http://sape.inf.usi.ch/quick-reference/ggpl
            size=1,
            alpha=0.5) +
  geom_point(aes(y = Hispanic.Prevalence, colour = 'Hispanic'),
             size=2,
             shape=20,
             alpha=0.5) +
  # Add line for Non.hispanic.black.Prevalence
  geom_line(aes(y = Non.hispanic.black.Prevalence, colour = 'Non_Hispanic_Blac
            linetype = "solid",  # http://sape.inf.usi.ch/quick-reference/ggpl
            size=1,
            alpha=0.5) +
  geom_point(aes(y = Non.hispanic.black.Prevalence, colour = 'Non_Hispanic_Bla
             size=2,
             shape=20,
             alpha=0.5) +
  # Add line for Non.hispanic.white.Prevalence
  geom_line(aes(y = Non.hispanic.white.Prevalence, colour = 'Non_Hispanic_Whit
            linetype = "solid",  # http://sape.inf.usi.ch/quick-reference/ggpl
            size=1,
            alpha=0.5) +
  geom_point(aes(y = Non.hispanic.white.Prevalence, colour = 'Non_Hispanic_Whi
             size=2,
             shape=20,
             alpha=0.5) +
  scale_colour_manual(name="",
                      labels = c("ADDM Average",
                                 "Asian/Pacific Islander",
                                 "Hispanic",
                                 "Non-Hispanic Black",
                                 "Non-Hispanic White"),
                      values = c(ADDM_Average="darkslategrey",
                                 Asian_Pacific_Islander ="darkred",
                                 Hispanic ="darkorchid3",
```

```
                                  Non_Hispanic_Black ="deepskyblue3",
                                  Non_Hispanic_White ="chartreuse3"))
# Add title, axis label, and axis scale
p <- p + scale_y_continuous(name = "Prevalence per 1,000 Children",
                            breaks = seq(5, 20, 5),
                            limits=c(5, 20)) +
  scale_x_continuous(name = "Year",
                     breaks = seq(2000, 2016, 1),
                     limits = c(2000, 2016)) +
  ggtitle("Prevalence Estimates by Race/Ethnicity [ Source: ADDM ]") +
  theme(title = element_text(face = 'bold.italic', color = "darkslategrey"),
        axis.title = element_text(face = 'plain', color = "darkslategrey"))
# Show plot
p
```

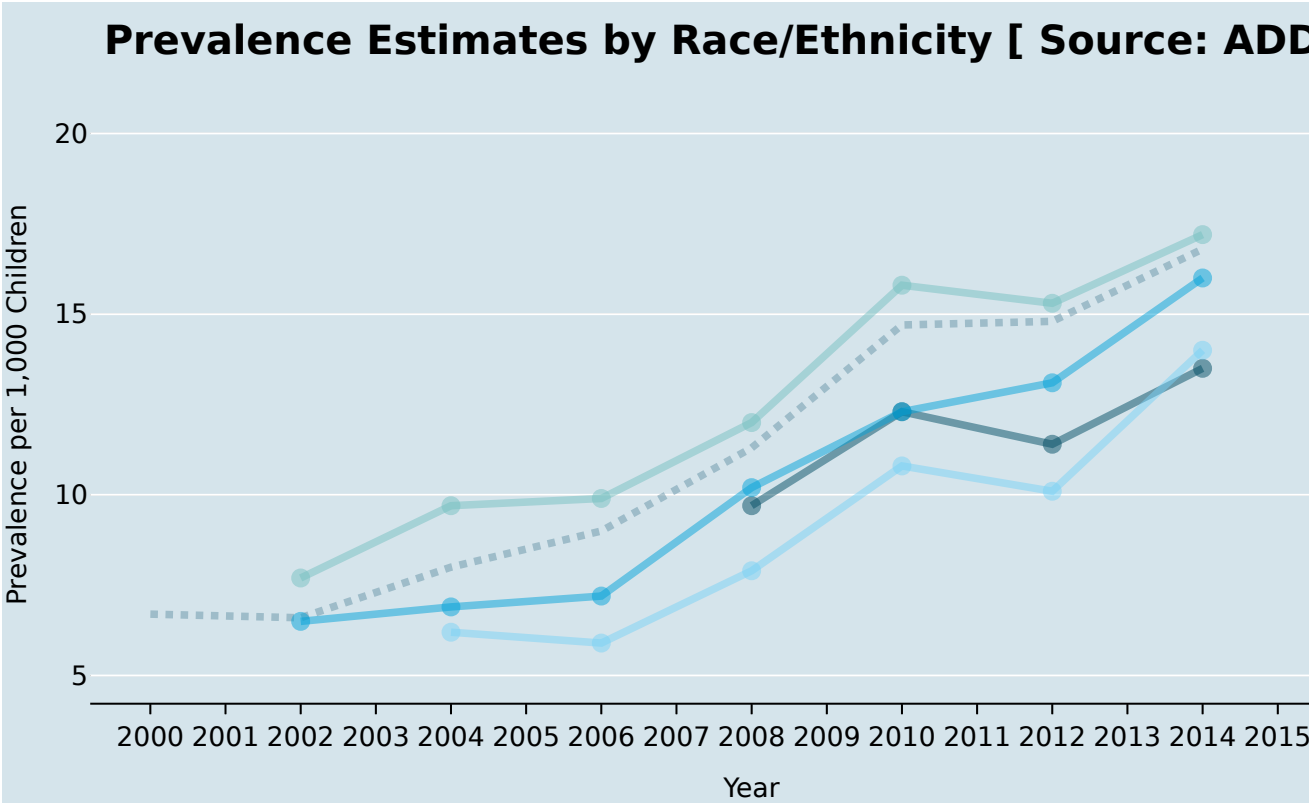
```
Warning message:
“Removed 4 rows containing missing values (geom_path).”Warning message:
“Removed 4 rows containing missing values (geom_point).”Warning message:
“Removed 2 rows containing missing values (geom_path).”Warning message:
“Removed 2 rows containing missing values (geom_point).”Warning message:
“Removed 1 rows containing missing values (geom_path).”Warning message:
“Removed 1 rows containing missing values (geom_point).”Warning message:
“Removed 1 rows containing missing values (geom_path).”Warning message:
“Removed 1 rows containing missing values (geom_point).”
```



In [80]:
```
# Apply theme
# p + theme_economist() + scale_colour_economist() # p + theme_wsj() + scale_c
```

In [81]:
```
# Dynamic chart:
p_dynamic <- p + theme_economist() + scale_colour_economist()
p_dynamic <- ggplotly(p_dynamic)
p_dynamic
```

```
Scale for 'colour' is already present. Adding another scale for 'colour',
which will replace the existing scale.
```



Quiz:

Change above zig-zag lines to spline/smooth lines.

Hints: Refer to ADDM Network estimates for overall ASD prevalence in US over time.

In [82]:
```
# Write your code below and press Shift+Enter to execute
```

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

()

Data Visualisation (Enhanced) - US. State Level Data Processing

In [83]:
```
# ---------------------------------
# Dataset: US. State Level Children ASD Prevalence
# ---------------------------------

ASD_State    <- read.csv("../dataset/ADV_ASD_State.csv", stringsAsFactors = FA

# Obtain number of rows and number of columns/features/variables
dim(ASD_State)
# Obtain overview (data structure/types)
str(ASD_State)
```

```
1692  49


'data.frame':   1692 obs. of  49 variables:
 $ State                                 : chr  "AZ" "GA" "MD" "NJ" ...
 $ Denominator                           : int  45322 43593 21532 29714 24535
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 2000
2002 2002 2002 2002 ...
 $ Source                                : chr  "addm" "addm" "addm" "addm"
...
 $ Source_Full1                          : chr  "Autism & Developmental Disab
ilities Monitoring Network" "Autism & Developmental Disabilities Monitoring N
etwork" "Autism & Developmental Disabilities Monitoring Network" "Autism & De
velopmental Disabilities Monitoring Network" ...
 $ State_Full1                           : chr  "Arizona" "Georgia" "Marylan
d" "New Jersey" ...
 $ State_Full2                           : chr  "AZ-Arizona" "GA-Georgia" "MD
-Maryland" "NJ-New Jersey" ...
 $ Numerator_ASD                         : int  295 283 118 294 155 104 117 2
80 252 65 ...
 $ Numerator_NonASD                      : int  45027 43310 21414 29420 24380
22961 35355 44833 36220 10955 ...
 $ Proportion                            : num  0.00651 0.00649 0.00548 0.009
89 0.00632 ...
 $ X95_Z_CI                              : num  0.00074 0.000754 0.000986 0.0
01125 0.000991 ...
 $ Z_Lower.CI                            : num  5.77 5.74 4.49 8.77 5.33 ...
 $ Z_Upper.CI                            : num  7.25 7.25 6.47 11.02 7.31 ...
 $ Z_Lower.CI_ABSerror                   : num  0.0314 0.062 0.1059 0.1311 0.
0739 ...
 $ Z_Upper.CI_ABSerror                   : num  0.0507 0.0542 0.1337 0.0803
0.0911 ...
 $ Chi_Wilson_P                          : num  0.00655 0.00654 0.00557 0.009
96 0.00639 ...
 $ X95_Chi_Wilson_CI                     : num  0.000741 0.000755 0.00099 0.0
01127 0.000994 ...
 $ Chi_Wilson_Lower.CI                   : num  5.81 5.78 4.58 8.83 5.4 ...
 $ Chi_Wilson_Upper.CI                   : num  7.29 7.29 6.56 11.08 7.39 ...
 $ Chi_Wilson_Lower.CI_ABSerror          : num  0.009314 0.019761 0.021503 0.
069416 0.000453 ...
 $ Chi_Wilson_Upper.CI_ABSerror          : num  0.0077 0.00953 0.04165 0.0152
3 0.01087 ...
 $ Chi_Wilson_Corrected_w_minus.CI       : num  0.0058 0.00577 0.00456 0.0088
1 0.00538 ...
 $ Chi_Wilson_Corrected_w_plus.CI        : num  0.0073 0.0073 0.00658 0.0111
0.00741 ...
 $ Chi_Wilson_Corrected_Lower.CI         : num  5.8 5.77 4.56 8.81 5.38 ...
 $ Chi_Wilson_Corrected_Upper.CI         : num  7.3 7.3 6.58 11.1 7.41 ...
```

```
 $ Chi_Wilson_Corrected_Lower.CI_ABSerror: num  0.00109 0.03057 0.04265 0.085
29 0.01834 ...
 $ Chi_Wilson_Corrected_Upper.CI_ABSerror: num  0.00395 0.0026 0.01636 0.0025
4 0.01108 ...
 $ Male.Prevalence                       : num  9.7 11 8.6 14.8 9.3 6.6 5 10.
1 10.7 9.9 ...
 $ Male.Lower.CI                         : num  8.5 9.7 7.1 13 7.8 5.2 4.1 8.
8 9.3 7.6 ...
 $ Male.Upper.CI                         : num  11.1 12.4 10.6 16.8 11.2 8.2
6.2 11.4 12.3 12.9 ...
 $ Female.Prevalence                     : num  3.2 2 2.2 4.3 3.3 2.4 1.4 2.2
2.9 1.7 ...
 $ Female.Lower.CI                       : num  2.5 1.5 1.5 3.3 2.4 1.6 0.9
1.7 2.2 0.9 ...
 $ Female.Upper.CI                       : num  4 2.7 2.7 5.5 4.5 3.5 2.1 2.9
3.8 3.2 ...
 $ Non.hispanic.white.Prevalence         : num  8.6 7.9 4.9 11.3 6.5 4.5 3.3
7.7 7.4 6.4 ...
 $ Non.hispanic.white.Lower.CI           : num  7.5 6.7 3.8 9.5 5.2 3.7 2.6
6.7 6.5 4.8 ...
 $ Non.hispanic.white.Upper.CI           : num  9.8 9.3 6.4 13.3 8.2 5.5 4.1
8.9 8.6 8.5 ...
 $ Non.hispanic.black.Prevalence         : chr  "7.3" "5.3" "6.1" "10.6" ...
 $ Non.hispanic.black.Lower.CI           : chr  "4.4" "4.4" "4.7" "8.5" ...
 $ Non.hispanic.black.Upper.CI           : chr  "12.2" "6.4" "8" "13.1" ...
 $ Hispanic.Prevalence                   : chr  "No data" "No data" "No data"
"No data" ...
 $ Hispanic.Lower.CI                     : chr  "No data" "No data" "No data"
"No data" ...
 $ Hispanic.Upper.CI                     : chr  "No data" "No data" "No data"
"No data" ...
 $ 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" ...
 $ State_Region                          : chr  "D8 Mountain" "D5 South Atlan
tic" "D5 South Atlantic" "D2 Middle Atlantic" ...
```

Data Visualisation (Enhanced) - US. State Level Data Pre-Process data

Pre-Process data: Missing data

In [84]:
```r
# 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
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 [85]:
```r
# Count missing values in dataframe:
sum(is.na(ASD_State)) # missing data recognised by R (NA)
# Define several offending strings
na_strings <- c("", "No data", "NA", "N A", "N / A", "N/A", "N/ A", "Not Avail
# Replace these defined missing values to R's internal NA
ASD_State = replace_with_na_all(ASD_State, condition = ~.x %in% na_strings)
# Count missing values in dataframe:
sum(is.na(ASD_State))
```

```
14454
```

```
28992
```

Remove invalid unicode char/string: \x92

In [86]:
```r
# Remove invalid unicode char/string: \x92
ASD_State$Source_Full1[ASD_State$Source_Full1 == "National Survey of Children\
```

Delete/Drop variable by index: column from 14 to 26, 29, and 30

In [87]:
```
cbind(names(ASD_State), c(1:length(names(ASD_State))))
```

State	1
Denominator	2
Prevalence	3
Lower.CI	4
Upper.CI	5
Year	6
Source	7
Source_Full1	8
State_Full1	9
State_Full2	10
Numerator_ASD	11
Numerator_NonASD	12
Proportion	13
X95_Z_CI	14
Z_Lower.CI	15
Z_Upper.CI	16
Z_Lower.CI_ABSerror	17
Z_Upper.CI_ABSerror	18
Chi_Wilson_P	19
X95_Chi_Wilson_CI	20
Chi_Wilson_Lower.CI	21
Chi_Wilson_Upper.CI	22
Chi_Wilson_Lower.CI_ABSerror	23
Chi_Wilson_Upper.CI_ABSerror	24
Chi_Wilson_Corrected_w_minus.CI	25
Chi_Wilson_Corrected_w_plus.CI	26
Chi_Wilson_Corrected_Lower.CI	27
Chi_Wilson_Corrected_Upper.CI	28
Chi_Wilson_Corrected_Lower.CI_ABSerror	29
Chi_Wilson_Corrected_Upper.CI_ABSerror	30
Male.Prevalence	31
Male.Lower.CI	32
Male.Upper.CI	33
Female.Prevalence	34
Female.Lower.CI	35
Female.Upper.CI	36
Non.hispanic.white.Prevalence	37
Non.hispanic.white.Lower.CI	38
Non.hispanic.white.Upper.CI	39
Non.hispanic.black.Prevalence	40
Non.hispanic.black.Lower.CI	41

Non.hispanic.black.Upper.CI	42
Hispanic.Prevalence	43
Hispanic.Lower.CI	44
Hispanic.Upper.CI	45
Asian.or.Pacific.Islander.Prevalence	46
Asian.or.Pacific.Islander.Lower.CI	47
Asian.or.Pacific.Islander.Upper.CI	48
State_Region	49

In [88]:
```
# Delete/Drop variable by index: column from 14 to 26, 29, and 30
# names(ASD_State)
ASD_State <- ASD_State[ -c(14:26, 29, 30) ]
```

Create new variables

In [89]:
```
# Create one new variable: Source_UC as uppercase of Source
ASD_State$Source_UC <- paste(toupper(ASD_State$Source))
# Create one new variable: Source_Full3 by combining Source_UC and Source_Full
ASD_State$Source_Full3 <- paste(ASD_State$Source_UC, ASD_State$Source_Full1)
```

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

In [90]:
```
# Recode Risk into category from Prevalence

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

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

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

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

In [91]:
```
# Recode Risk into category from Prevalence

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

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

Convert to correct data types

In [92]:
```
str(ASD_State)
```

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       1692 obs. of  38 variables:
 $ State                                : chr  "AZ" "GA" "MD" "NJ" ...
 $ Denominator                          : int  45322 43593 21532 29714 24535 2
3065 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 2000 2
002 2002 2002 2002 ...
 $ Source                               : chr  "addm" "addm" "addm" "addm" ...
 $ 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" ...
 $ State_Full1                          : chr  "Arizona" "Georgia" "Maryland"
"New Jersey" ...
 $ State_Full2                          : chr  "AZ-Arizona" "GA-Georgia" "MD-M
aryland" "NJ-New Jersey" ...
 $ Numerator_ASD                        : int  295 283 118 294 155 104 117 280
252 65 ...
 $ Numerator_NonASD                     : int  45027 43310 21414 29420 24380 2
2961 35355 44833 36220 10955 ...
 $ Proportion                           : num  0.00651 0.00649 0.00548 0.00989
0.00632 ...
 $ Chi_Wilson_Corrected_Lower.CI        : num  5.8 5.77 4.56 8.81 5.38 ...
 $ Chi_Wilson_Corrected_Upper.CI        : num  7.3 7.3 6.58 11.1 7.41 ...
 $ Male.Prevalence                      : num  9.7 11 8.6 14.8 9.3 6.6 5 10.1
10.7 9.9 ...
 $ Male.Lower.CI                        : num  8.5 9.7 7.1 13 7.8 5.2 4.1 8.8
9.3 7.6 ...
 $ Male.Upper.CI                        : num  11.1 12.4 10.6 16.8 11.2 8.2 6.
2 11.4 12.3 12.9 ...
 $ Female.Prevalence                    : num  3.2 2 2.2 4.3 3.3 2.4 1.4 2.2
2.9 1.7 ...
 $ Female.Lower.CI                      : num  2.5 1.5 1.5 3.3 2.4 1.6 0.9 1.7
2.2 0.9 ...
 $ Female.Upper.CI                      : num  4 2.7 2.7 5.5 4.5 3.5 2.1 2.9
3.8 3.2 ...
 $ Non.hispanic.white.Prevalence        : num  8.6 7.9 4.9 11.3 6.5 4.5 3.3 7.
7 7.4 6.4 ...
 $ Non.hispanic.white.Lower.CI          : num  7.5 6.7 3.8 9.5 5.2 3.7 2.6 6.7
6.5 4.8 ...
 $ Non.hispanic.white.Upper.CI          : num  9.8 9.3 6.4 13.3 8.2 5.5 4.1 8.
9 8.6 8.5 ...
 $ Non.hispanic.black.Prevalence        : chr  "7.3" "5.3" "6.1" "10.6" ...
 $ Non.hispanic.black.Lower.CI          : chr  "4.4" "4.4" "4.7" "8.5" ...
 $ Non.hispanic.black.Upper.CI          : chr  "12.2" "6.4" "8" "13.1" ...
 $ Hispanic.Prevalence                  : chr  NA NA NA NA ...
 $ Hispanic.Lower.CI                    : chr  NA NA NA NA ...
 $ Hispanic.Upper.CI                    : chr  NA NA NA NA ...
 $ 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 ...
 $ State_Region                         : chr  "D8 Mountain" "D5 South Atlanti
c" "D5 South Atlantic" "D2 Middle Atlantic" ...
 $ 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" ...
```

In [93]:
```
# cbind(names(ASD_State), c(1:length(names(ASD_State))))
```

Convert variables to numeric

In [94]:
```
# Convert Prevalence and CIs from categorical/chr to numeric
ix <- 13:33 # define an index
ASD_State[ix] <- lapply(ASD_State[ix], as.numeric)
```

Convert variables to categorical/factor

In [95]:
```
# Convert Source from categorical/chr to categorical/factor
ix <- c(1, 7, 8, 9, 10, 34, 35, 36) # define an index
ASD_State[ix] <- lapply(ASD_State[ix], as.factor)

# Create new ordered factor Year_Factor from Year
ASD_State$Year_Factor <- factor(ASD_State$Year, ordered = TRUE)
```

Convert Prevalence_Rank2 & Prevalence_Rank4 to ordered factor

In [96]:
```
# Convert to factor
ASD_State$Prevalence_Risk2 = factor(ASD_State$Prevalence_Risk2, ordered=TRUE,
                                    levels=c("Low", "High"))
# Convert to factor
ASD_State$Prevalence_Risk4 = factor(ASD_State$Prevalence_Risk4, ordered=TRUE,
                                    levels=c("Low", "Medium", "High", "
```

In [97]:
```
# Display unique values (levels) of a factor categrotical
lapply(select_if(ASD_State, is.factor), levels)
```

$State
'AK' 'AL' 'AR' 'AZ' 'CA' 'CO' 'CT' 'DC' 'DE' 'FL' 'GA' 'HI' 'IA' 'ID' 'IL' 'IN' 'KS' 'KY' 'LA' 'MA' 'MD' 'ME' 'MI' 'MN' 'MO' 'MS' 'MT' 'NC' 'ND' 'NE' 'NH' 'NJ' 'NM' 'NV' 'NY' 'OH' 'OK' 'OR' 'PA' 'RI' 'SC' 'SD' 'TN' 'TX' 'UT' 'VA' 'VT' 'WA' 'WI' 'WV' 'WY'

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

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

$State_Full1
'Alabama' 'Alaska' 'Arizona' 'Arkansas' 'California' 'Colorado' 'Connecticut' 'Delaware' 'District of Columbia' 'Florida' 'Georgia' 'Hawaii' 'Idaho' 'Illinois' 'Indiana' 'Iowa' 'Kansas' 'Kentucky' 'Louisiana' 'Maine' 'Maryland' 'Massachusetts' 'Michigan' 'Minnesota' 'Mississippi' 'Missouri' 'Montana' 'Nebraska' 'Nevada' 'New Hampshire' 'New Jersey' 'New Mexico' 'New York' 'North Carolina' 'North Dakota' 'Ohio' 'Oklahoma' 'Oregon' 'Pennsylvania' 'Rhode Island' 'South Carolina' 'South Dakota' 'Tennessee' 'Texas' 'Utah' 'Vermont' 'Virginia' 'Washington' 'West Virginia' 'Wisconsin' 'Wyoming'

$State_Full2
'AK-Alaska' 'AL-Alabama' 'AR-Arkansas' 'AZ-Arizona' 'CA-California' 'CO-Colorado' 'CT-Connecticut' 'DC-District of Columbia' 'DE-Delaware' 'FL-Florida' 'GA-Georgia' 'HI-Hawaii' 'IA-Iowa' 'ID-Idaho' 'IL-Illinois' 'IN-Indiana' 'KS-Kansas' 'KY-Kentucky' 'LA-Louisiana' 'MA-Massachusetts' 'MD-Maryland' 'ME-Maine' 'MI-Michigan' 'MN-Minnesota' 'MO-Missouri' 'MS-Mississippi' 'MT-Montana' 'NC-North Carolina' 'ND-North Dakota' 'NE-Nebraska' 'NH-New Hampshire' 'NJ-New Jersey' 'NM-New Mexico' 'NV-Nevada' 'NY-New York' 'OH-Ohio' 'OK-Oklahoma' 'OR-Oregon' 'PA-Pennsylvania' 'RI-Rhode Island' 'SC-South Carolina' 'SD-South Dakota' 'TN-Tennessee' 'TX-Texas' 'UT-Utah' 'VA-Virginia' 'VT-Vermont' 'WA-Washington' 'WI-Wisconsin' 'WV-West Virginia' 'WY-Wyoming'

$State_Region
'D1 New England' 'D2 Middle Atlantic' 'D3 East North Central' 'D4 West North Central' 'D5 South Atlantic' 'D6 East South Central' 'D7 West South Central' 'D8 Mountain' 'D9 Pacific'

$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'

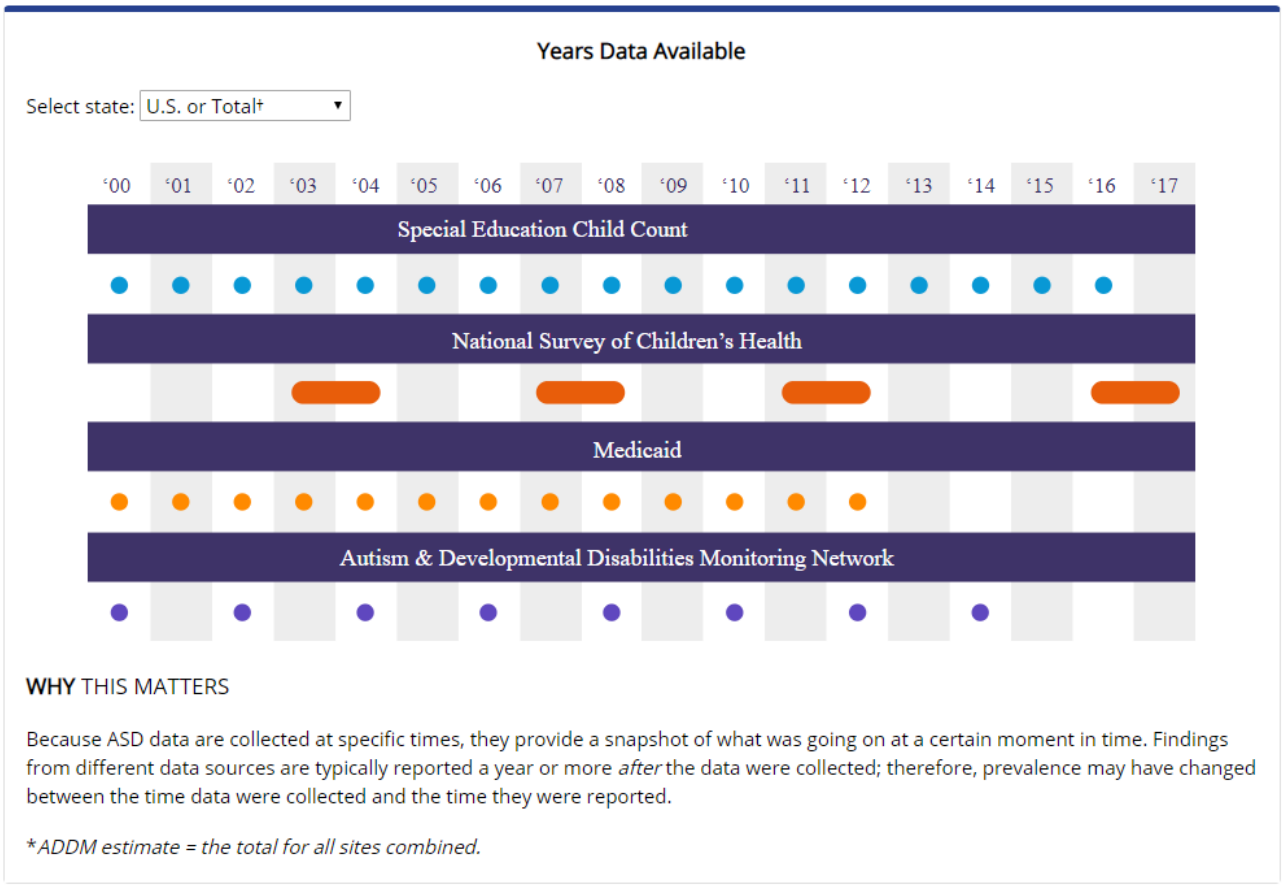
$Year_Factor
'2000' '2001' '2002' '2003' '2004' '2005' '2006' '2007' '2008' '2009' '2010' '2011' '2012' '2013' '2014' '2015' '2016'

Optionally, export the processed dataframe data to CSV file.

In [98]:
```
write.csv(ASD_State, file = "../dataset/ADV_ASD_State_R.csv", row.names = FALS
```

In [99]:
```
# Read back in above saved file:
# ASD_State <- read.csv("../dataset/ADV_ASD_State_R.csv")
# ASD_State$Year_Factor <- factor(ASD_State$Year_Factor, ordered = TRUE) # Con
# ASD_State$Prevalence_Risk2 = factor(ASD_State$Prevalence_Risk2, ordered=TRUE
# ASD_State$Prevalence_Risk4 = factor(ASD_State$Prevalence_Risk4, ordered=TRUE
```

Data Visualisation (Enhanced) - US. State Level Data Visualisation



Above chat shows at data source level, we'd also like to know State level data availbility. How?

Data Visualisation (Enhanced) - [R] Explore the Data [Years Data Available by State]

In [100]:
```
# Adjust in-line plot size to M x N
options(repr.plot.width=8, repr.plot.height=12)
```

In [101]:

```
# ---------------------------------
# [State] < Years Data Available by State >
# ---------------------------------
p <- ggplot(ASD_State, aes(x = Source, fill = Source)) +
  geom_bar() + theme(axis.text.x=element_blank(),  # Hide axis
                     axis.ticks.x=element_blank(), # Hide axis
                     axis.text.y=element_blank(),  # Hide axis
                     axis.ticks.y=element_blank(), # Hide axis
                     panel.background = element_blank(), # Remove panel backgr
                     legend.position="top",
                     strip.text.y = element_text(angle=0) # Rotate text to hor
  ) +
  scale_fill_manual("Data Source:", values = c("addm" = "darkblue",
                                               "medi" = "orange",
                                               "nsch" = "darkred",
                                               "sped" = "skyblue")) +
  facet_grid(facets = State_Full2 ~ Year) +
  labs(x="", y="", title="Years Data Available by State") # layers of graphics
```

In [102]:
```
# Below plot may run for a while
# Show plot
p
```

Filter and create dataframe of different data sources, for easy data access

In [103]:
```
# Filter and create dataframe of different data sources, for easy data access
ASD_State_ADDM <- subset(ASD_State, Source == 'addm')
ASD_State_MEDI <- subset(ASD_State, Source == 'medi')
ASD_State_NSCH <- subset(ASD_State, Source == 'nsch')
ASD_State_SPED <- subset(ASD_State, Source == 'sped')
```

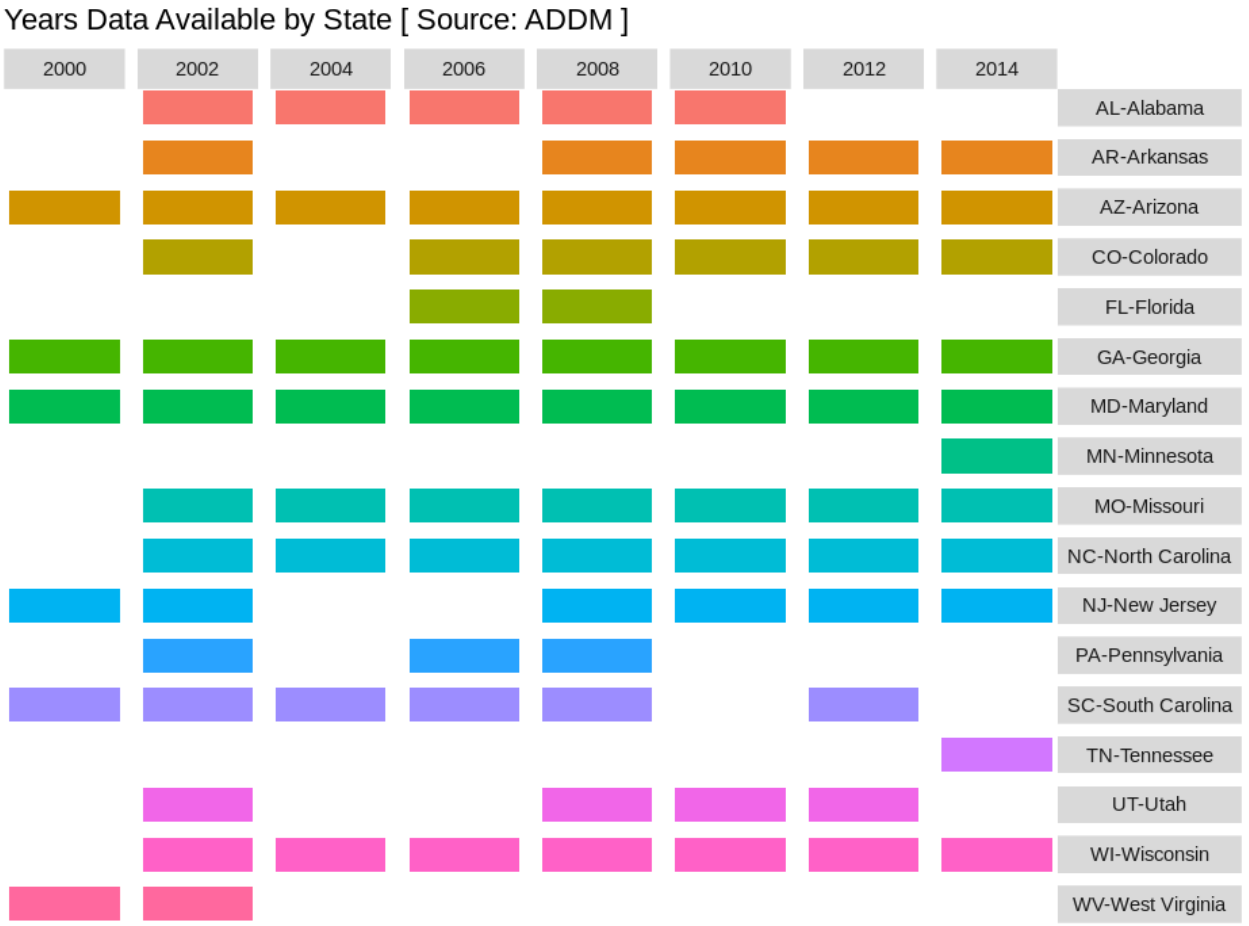
Data Visualisation (Enhanced) - [R] Explore the Data Years Data Available by State [Source: ADDM]

In [104]:
```
# Adjust in-line plot size to M x N
options(repr.plot.width=8, repr.plot.height=6)
```

Years Data Available by State [Source: ADDM]

In [105]:
```
# Years Data Available by State [ Source: ADDM ]
p <- ggplot(ASD_State_ADDM, aes(x = 1, fill = State_Full2)) +
  geom_bar() + theme(axis.text.x=element_blank(),  # Hide axis
                     axis.ticks.x=element_blank(), # Hide axis
                     axis.text.y=element_blank(),  # Hide axis
                     axis.ticks.y=element_blank(), # Hide axis
                     panel.background = element_blank(), # Remove panel backgr
                     legend.position="none",
                     strip.text.y = element_text(angle=0) # Rotate text to hor
  ) +
  facet_grid(facets = State_Full2 ~ Year_Factor) +
  labs(x="", y="", title="Years Data Available by State [ Source: ADDM ]") # l
```

In [106]:
```
# Show plot
p
```



Quiz:

Create Years Data Available by State [Source: XXXX] for other three data sources:

In [107]:
```
# Write your code below and press Shift+Enter to execute
```

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

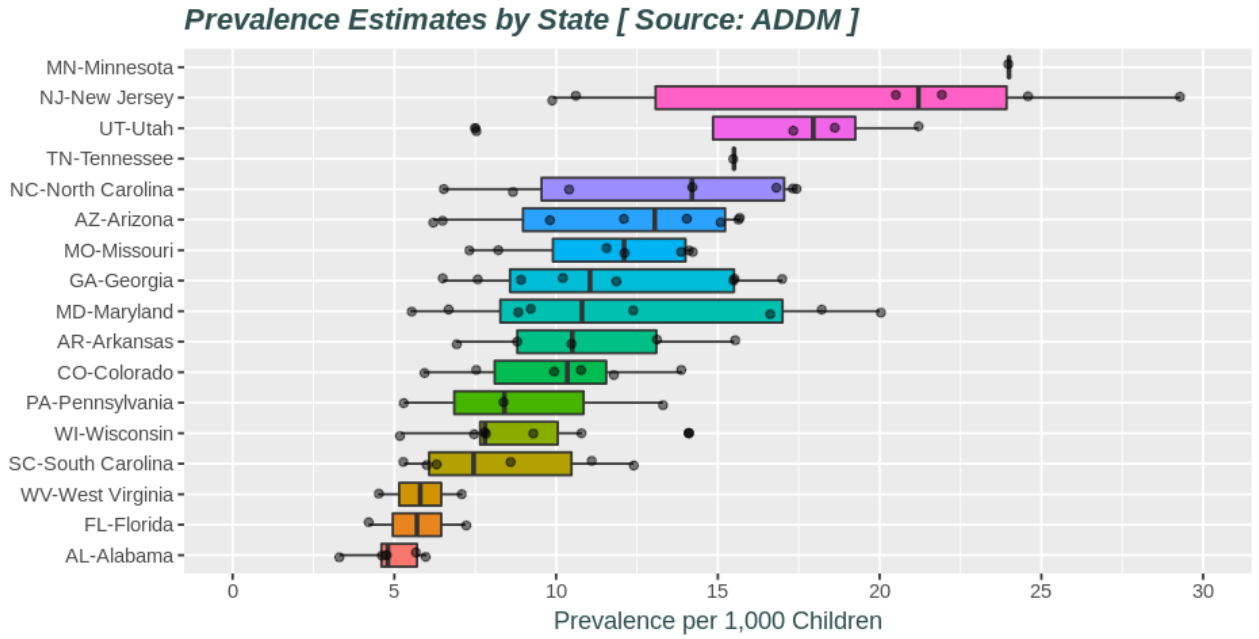
Data Visualisation (Enhanced) - [R] REPORTED PREVALENCE VARIES BY GEOGRAPHIC LOCATION (States) Prevalence Estimates by State [Source: ADDM]

In [108]:
```
# Adjust in-line plot size to M x N
options(repr.plot.width=8, repr.plot.height=4)
```

Visualise: **Prevalence Estimates by State [Source: ADDM]**

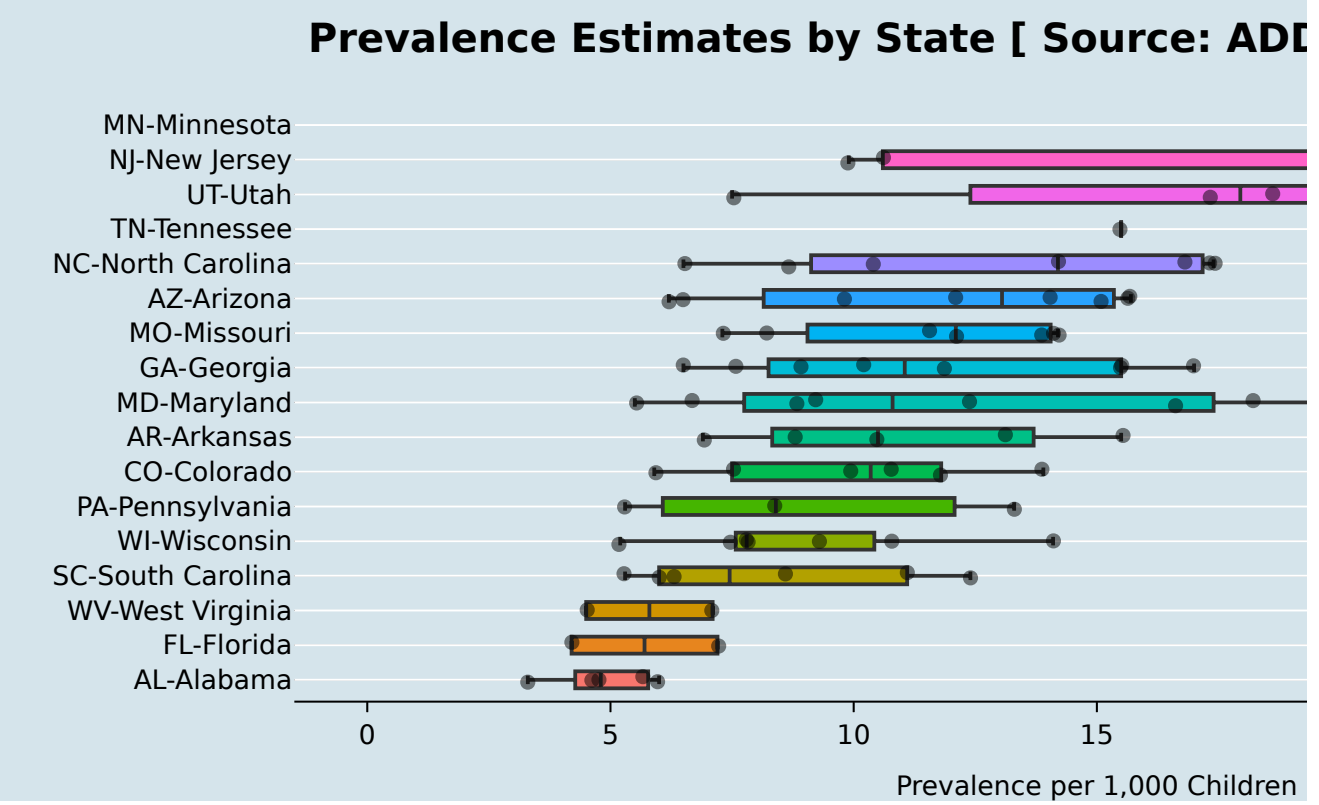
In [109]:
```
# Prevalence Estimates by State [ Source: ADDM ] , aggregated for different ye
p <- ggplot(ASD_State_ADDM, aes(x = reorder(State_Full2, Prevalence, FUN = med
                                y = Prevalence)) +
  geom_boxplot(aes(fill = reorder(State_Full2, Prevalence, FUN = median))) + #
  scale_fill_discrete(guide = guide_legend(title = "US. States")) + # Legend N
  #  geom_boxplot(fill = 'darkslategrey', alpha = 0.2) +
  scale_y_continuous(name = "Prevalence per 1,000 Children",
                     breaks = seq(0, 30, 5),
                     limits=c(0, 30)) +
  scale_x_discrete(name = "") +
  ggtitle("Prevalence Estimates by State [ Source: ADDM ]") +
  theme(title = element_text(face = 'bold.italic', color = "darkslategrey"),
        axis.title = element_text(face = 'plain', color = "darkslategrey"),
        legend.position = 'none') +
  coord_flip() + # Rotate chart
  geom_jitter(alpha = 0.5, position = position_jitter(width = 0.1)) # Add actu
```

In [110]:
```
# Show plot
p
```



In [111]:
```
# Theme of the economist magazine:
# p + theme_economist() + scale_colour_economist() + theme(legend.position = '
```

In [112]:
```
# Dynamic chart
p_dynamic <- p + theme_economist() + scale_colour_economist() + theme(legend.p
p_dynamic <- ggplotly(p_dynamic)
p_dynamic
```



Quiz:

Create Prevalence Estimates by State [Source: XXXX] for other three data sources:

In [113]:
```
# Write your code below and press Shift+Enter to execute
```

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

Data Visualisation (Enhanced) - [R] US. State Level No. Children Surveyed by State [Source: ADDM] [Year 2014]

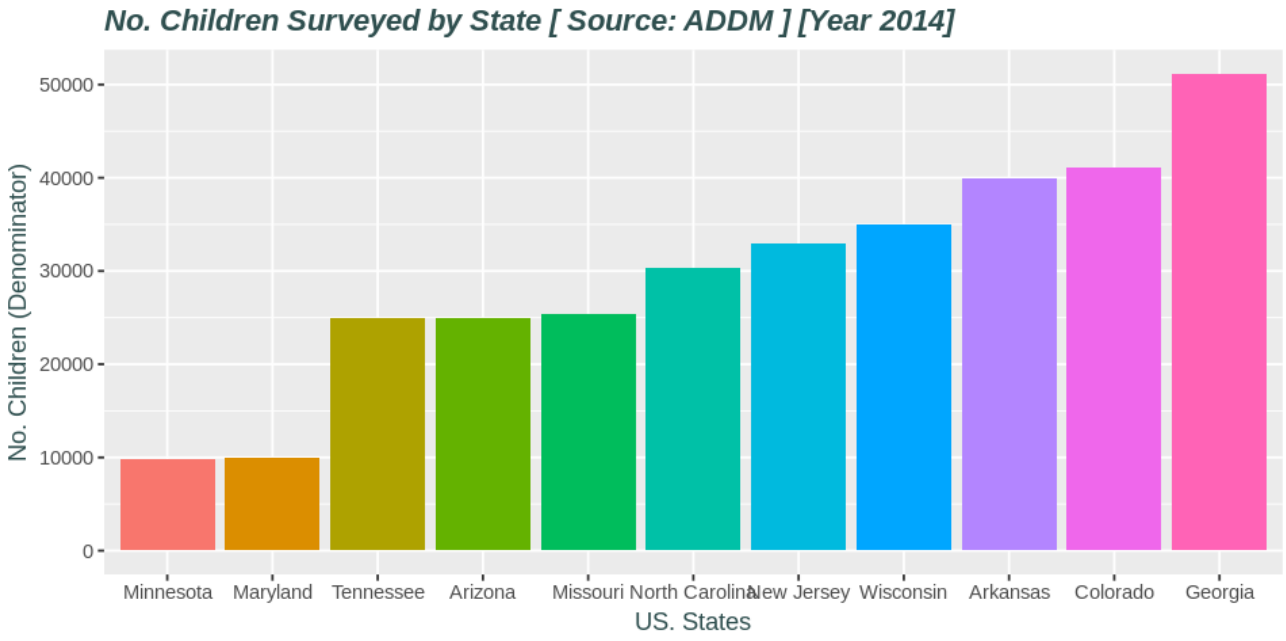
In [114]:
```
# Adjust in-line plot size to M x N
options(repr.plot.width=8, repr.plot.height=4)
```

Visualise: **No. Children Surveyed by State [Source: ADDM] [Year 2014]**

In [115]:
```
# All State Prevalence data with: Source == 'addm' & Year == 2014
# filter using dataframe: ASD_State_ADDM
ASD_State_Subset <- subset(ASD_State_ADDM, Year == 2014)
# or filer using dataframe: ASD_State
ASD_State_Subset <- subset(ASD_State, Source == 'addm' & Year == 2014)
```

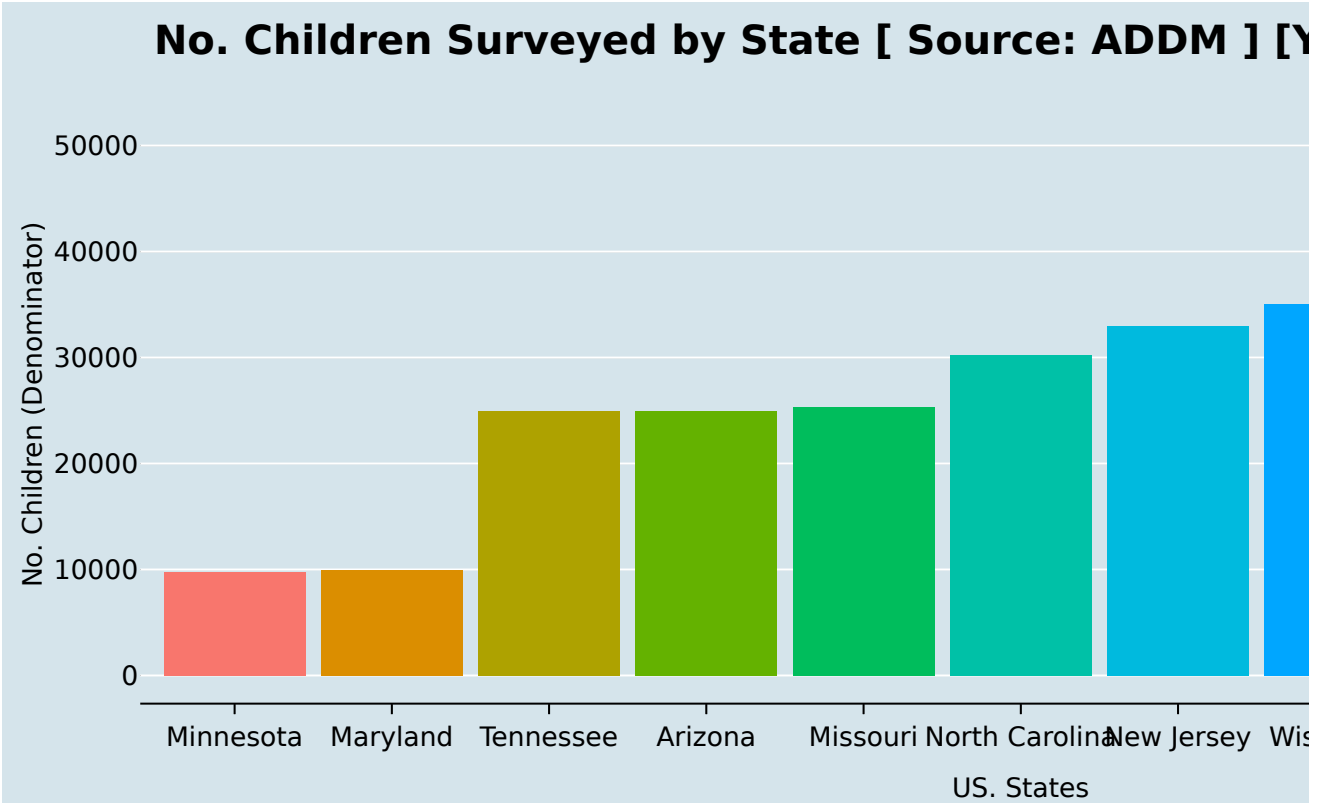
In [116]:
```
# Bar plot/chart for < No. Children surveyed by State [ADDM] [Year 2014] >
p <- ggplot(ASD_State_Subset, aes(x = reorder(State_Full1, Denominator, FUN =
                                  y = Denominator)) +
  geom_bar(stat="identity", aes(fill = reorder(State_Full1, Denominator, FUN =
  scale_fill_discrete(guide = guide_legend(title = "US. States")) + # Legend N
  scale_x_discrete(name = "US. States") +
  scale_y_continuous(name = "No. Children (Denominator)") +
  ggtitle("No. Children Surveyed by State [ Source: ADDM ] [Year 2014]") +
  #  geom_text(aes(label=Denominator), vjust=1.6, color="darkslategrey", size=
  theme(title = element_text(face = 'bold.italic', color = "darkslategrey"),
        axis.title = element_text(face = 'plain', color = "darkslategrey"),
        legend.position="none")
```

In [117]:
```
# Show plot
p
```



In [118]:
```
# Theme of the economist magazine:
# p + theme_economist() + scale_colour_economist() + theme(legend.position = '
```

In [119]:
```
# Dynamic chart
p_dynamic <- p + theme_economist() + scale_colour_economist() + theme(legend.p
p_dynamic <- ggplotly(p_dynamic)
p_dynamic
```



Quiz:

Create No. Children Surveyed by State [Source: XXXX] [Year CCYY] for other data sources & years:

In [120]:
```
# Write your code below and press Shift+Enter to execute
```

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

Quiz:

Create No. ASD Children by State [Source: XXXX] [Year CCYY] for other data sources & years:

Hint: Use variable: ASD_State_ADDM$Numerator_ASD

In [121]:
```
# Write your code below and press Shift+Enter to execute
```

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

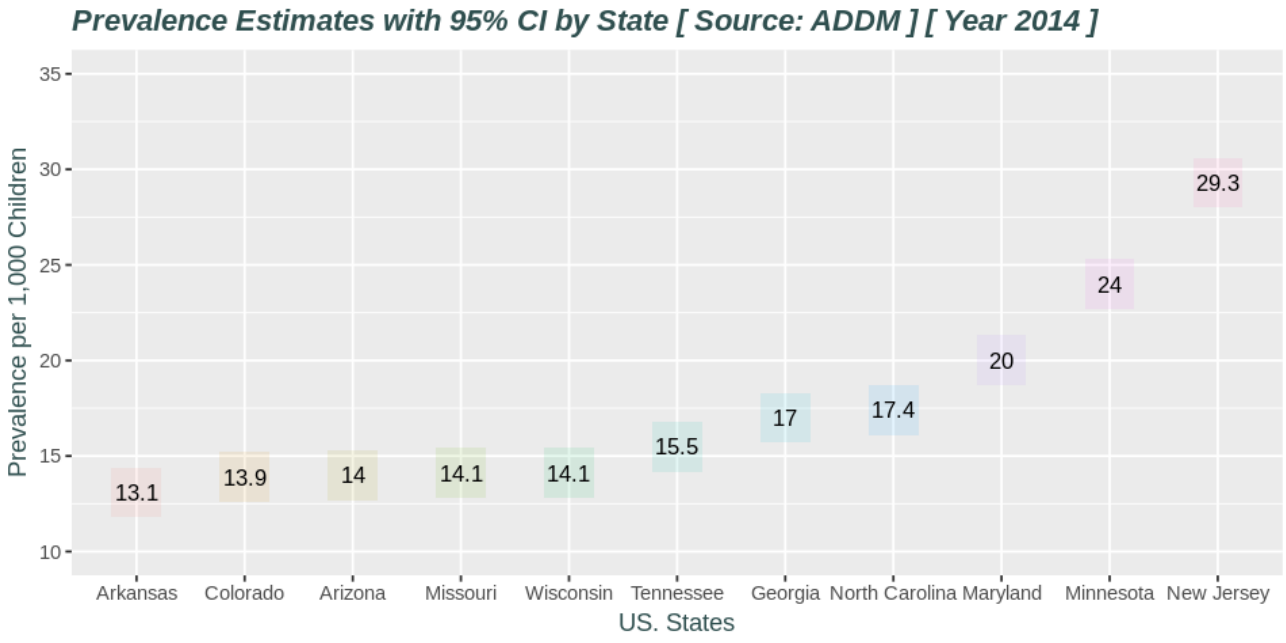
Data Visualisation (Enhanced) - [R] US. State Level Prevalence Estimates with 95% CI by State [Source: ADDM] [Year 2014]

In [122]:
```
# Adjust in-line plot size to M x N
# options(repr.plot.width=8, repr.plot.height=4)
```

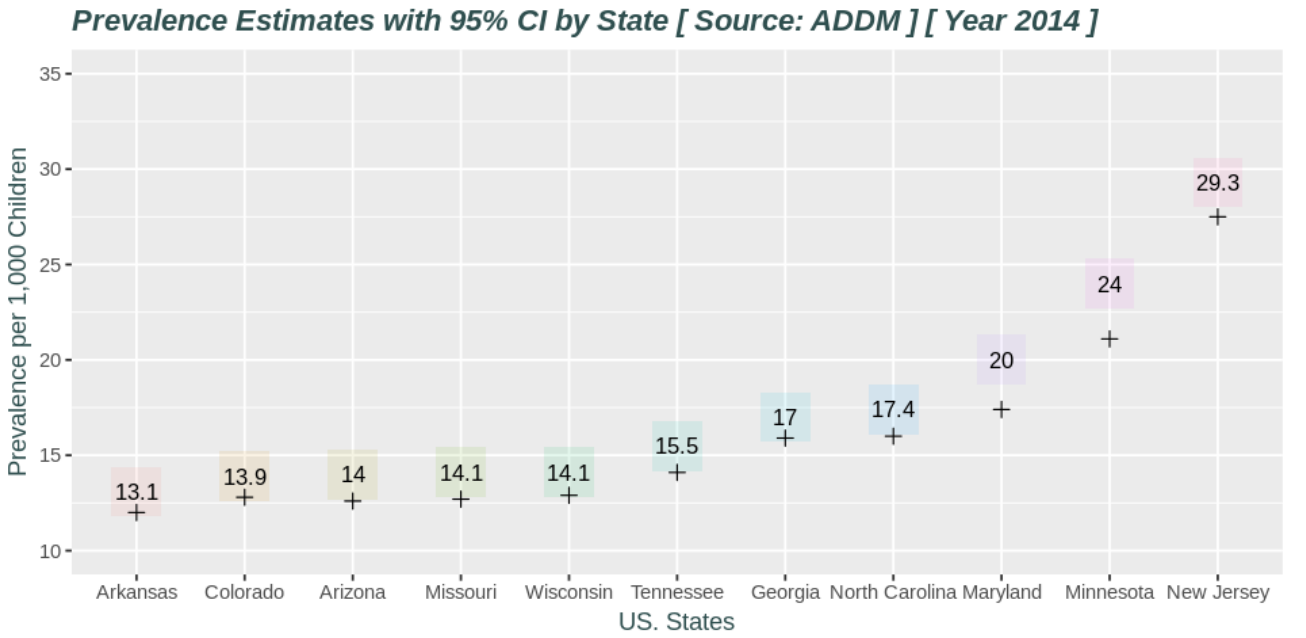
Visualise: **Prevalence Estimates with 95% CI by State [Source: ADDM] [Year 2014]**

In [123]:
```
# ASD_State_Subset <- subset(ASD_State_ADDM, Year == 2014)
# or
# ASD_State_Subset <- subset(ASD_State, Source == 'addm' & Year == 2014)

# Point plot/chart
p = ggplot(ASD_State_Subset, aes(x = reorder(State_Full1, Prevalence, median),
                                 y = Prevalence)) +
  geom_point(stat="identity", aes(colour = reorder(State_Full1, Prevalence, me
  scale_colour_discrete(guide = guide_legend(title = "US. States")) + # Legend
  scale_y_continuous(name = "Prevalence per 1,000 Children",
                     breaks = seq(10, 35, 5),
                     limits=c(10, 35)) +
  scale_x_discrete(name = "US. States") +
  ggtitle("Prevalence Estimates with 95% CI by State [ Source: ADDM ] [ Year 2
  theme(title = element_text(face = 'bold.italic', color = "darkslategrey"),
        axis.title = element_text(face = 'plain', color = "darkslategrey"),
        legend.position = 'none') +
  geom_text(aes(label=Prevalence), hjust=0.5, color="black", size=3.5)  # Show
```

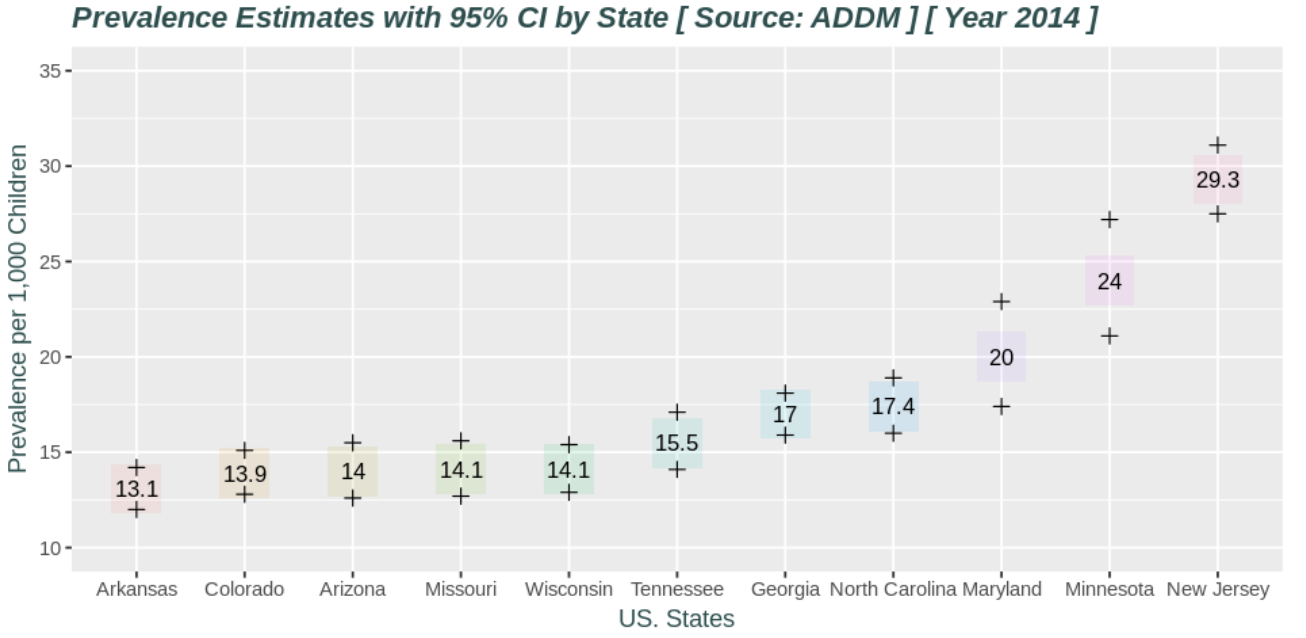
In [124]:
```
# Show plot
p
```



In [125]:
```
# Add Lower.CI
p = p + geom_point(data = ASD_State_Subset, aes(x = reorder(State_Full1, Preva
                                                shape=Source # point shape
),
size = 2 # point size
) +
  #  geom_text(aes(label=Lower.CI), hjust=-0.1, vjust=3, color="darkslategrey"
  scale_shape_manual(values=3)  # manual define point shape
# Show plot
p
```



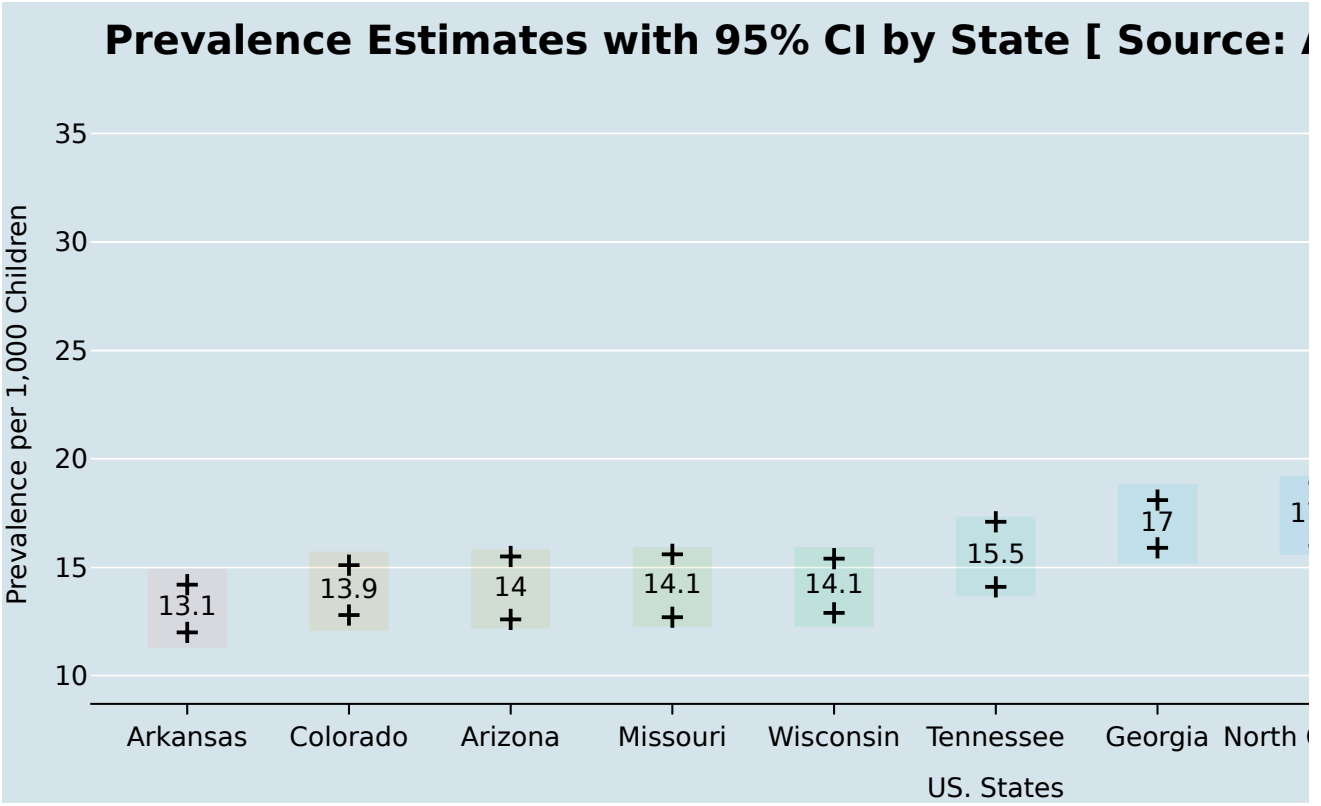
In [126]:
```
# Add Upper.CI
p = p + geom_point(data = ASD_State_Subset, aes(x = reorder(State_Full1, Preva
                                                shape=Source # point shape
),
size = 2 # point size
)
#  geom_text(aes(label=Upper.CI), hjust=-0.1, vjust=-3, color="darkslategrey",
# Show plot
p
```



In [127]:
```
# theme of the economist magazine:
# p + theme_economist() + scale_colour_economist() + scale_colour_discrete(gui
```

In [128]:
```
# Dynamic chart
p_dynamic <- p + theme_economist() + scale_colour_economist() + scale_colour_d
p_dynamic <- ggplotly(p_dynamic)
p_dynamic
```

```
Scale for 'colour' is already present. Adding another scale for 'colour',
which will replace the existing scale.
Scale for 'colour' is already present. Adding another scale for 'colour',
which will replace the existing scale.
```



Quiz:

Create Prevalence Estimates with 95% CI by State [Source: ADDM] [Year CCYY] for other data sources & years:

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

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

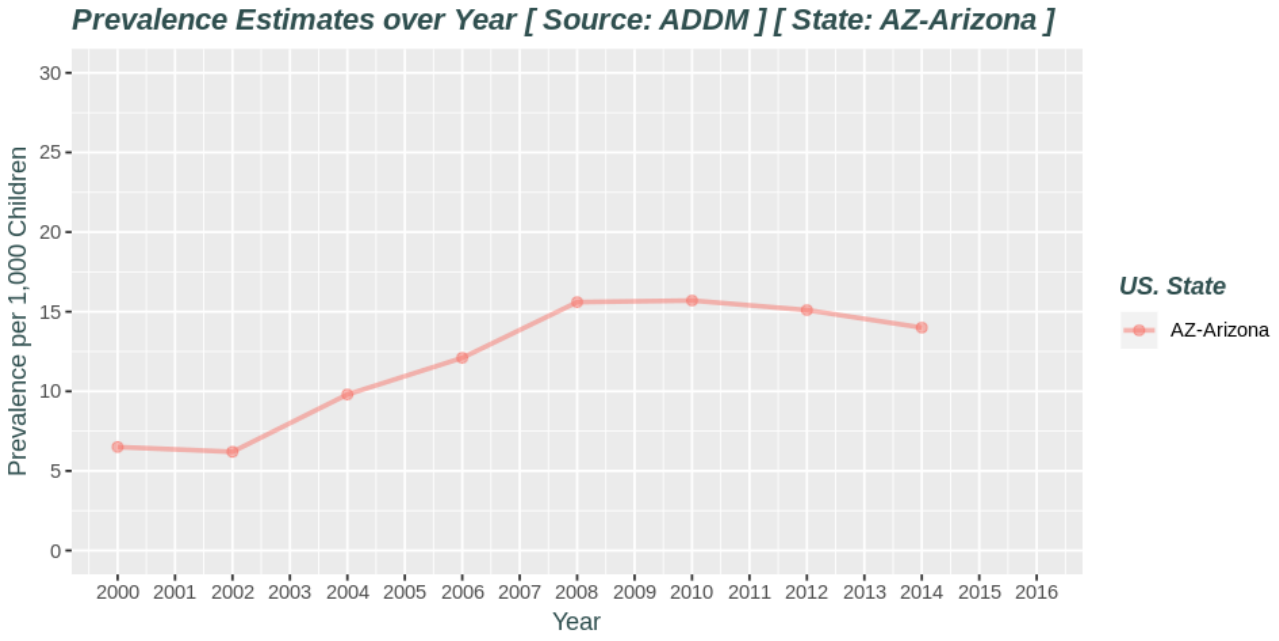
Data Visualisation (Enhanced) - [R] US. State Level Prevalence Estimates over Year [Source: ADDM] [State: AZ-Arizona]

In [130]:
```
# Adjust in-line plot size to M x N
# options(repr.plot.width=8, repr.plot.height=4)
```

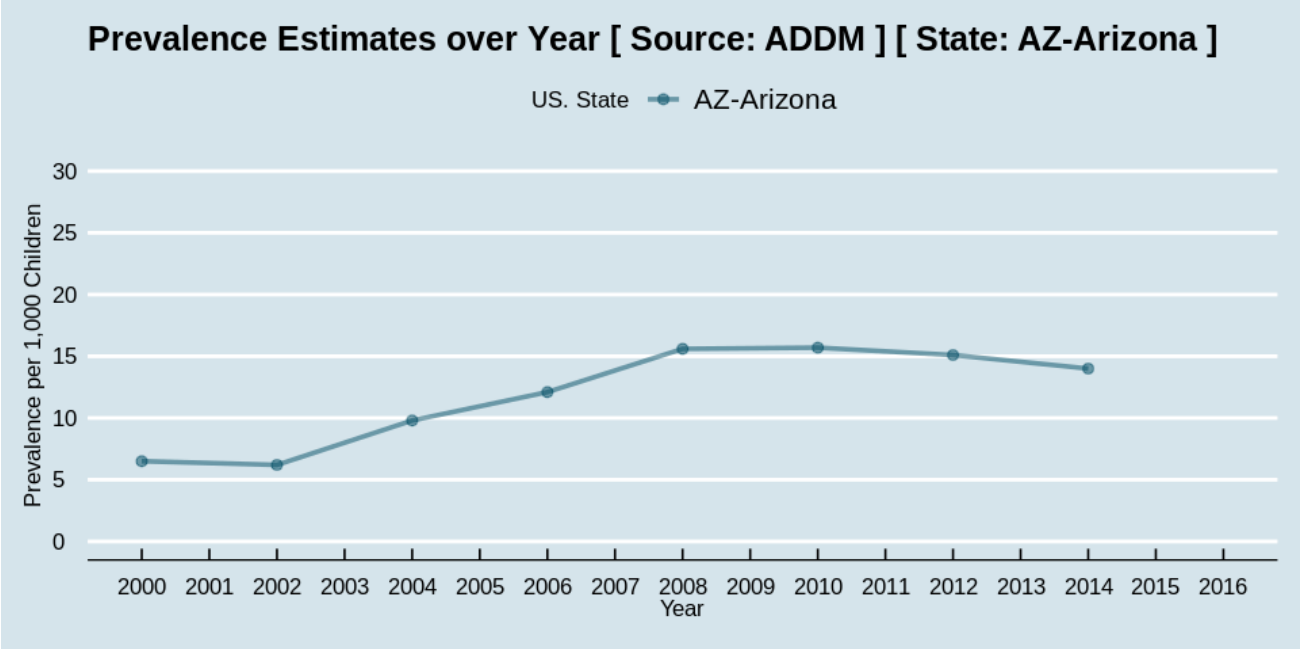
Visualise: **Prevalence Estimates over Year [Source: ADDM] [State: AZ-Arizona]**

In [131]:
```
# All year/time Prevalence data with: Source_UC == 'ADDM' & State_Full2 == 'AZ
ASD_State_Subset <- subset(ASD_State, Source_UC == 'ADDM' & State_Full2 == 'AZ

# Line plot/chart for < State ASD Prevalence [ADDM] [AZ-Arizona] >
p <- ggplot(ASD_State_Subset, aes(x = Year, y = Prevalence))
# Select (add) line chart type:
p <- p + geom_line(aes(color = State_Full2),
                   linetype = "solid",  # http://sape.inf.usi.ch/quick-referen
                   size=1,
                   alpha=0.5)
# Select (add) points to chart:
p <- p + geom_point(aes(color = State_Full2),
                    size=3,
                    shape=20,
                    alpha=0.5)
# Customize legend name:
p <- p + labs(color = "US. State")
# Adjust x and y axis, scale, limit and labels:
p <- p + scale_y_continuous(name = "Prevalence per 1,000 Children",
                            breaks = seq(0, 30, 5),
                            limits=c(0, 30)) +
  scale_x_continuous(name = "Year",
                     breaks = seq(2000, 2016, 1),
                     limits = c(2000, 2016))
# Customize chart title:
p <- p + ggtitle("Prevalence Estimates over Year [ Source: ADDM ] [ State: AZ-
# Customize chart title and axis labels:
p <- p + theme(title = element_text(face = 'bold.italic', color = "darkslategr
               axis.title = element_text(face = 'plain', color = "darkslategre
```

In [132]:
```
# Show plot
p
```



In [133]:
```
# Theme of the economist magazine:
p + theme_economist() + scale_colour_economist()
```



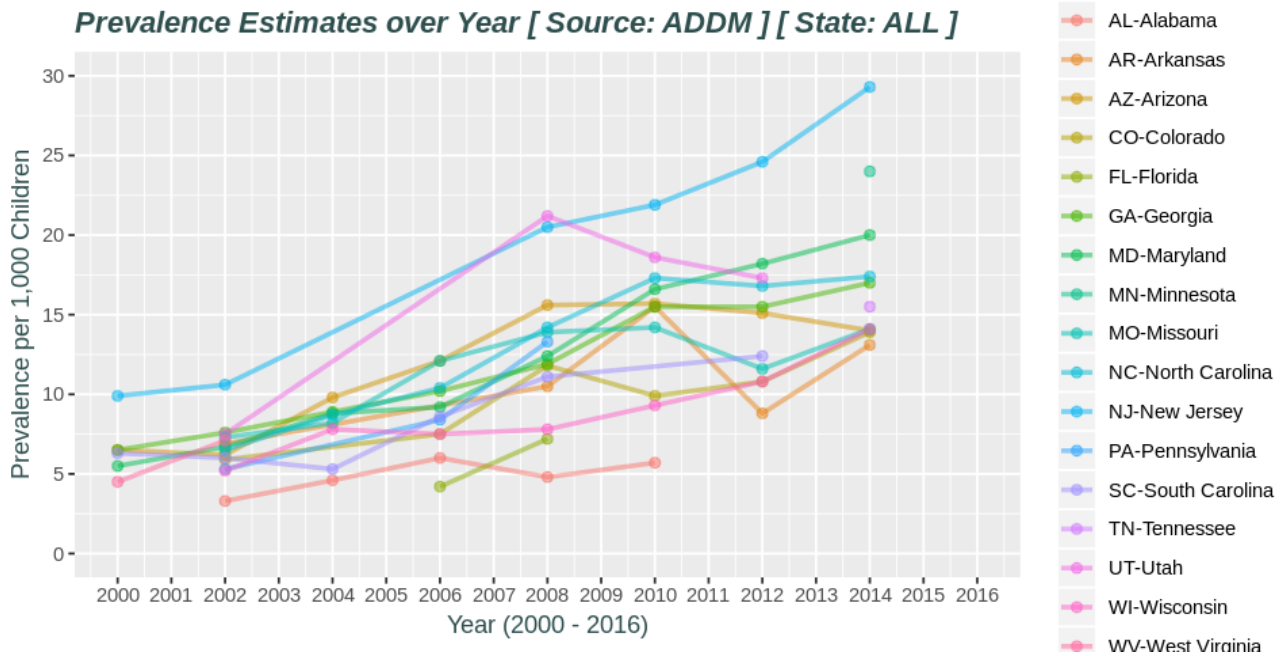
Data Visualisation (Enhanced) - [R] US. State Level Prevalence Estimates over Year [Source: ADDM] [State: ALL]

In [134]:
```
# Adjust in-line plot size to M x N
# options(repr.plot.width=8, repr.plot.height=4)
```

Visualise: **Prevalence Estimates over Year [Source: ADDM] [State: ALL]**

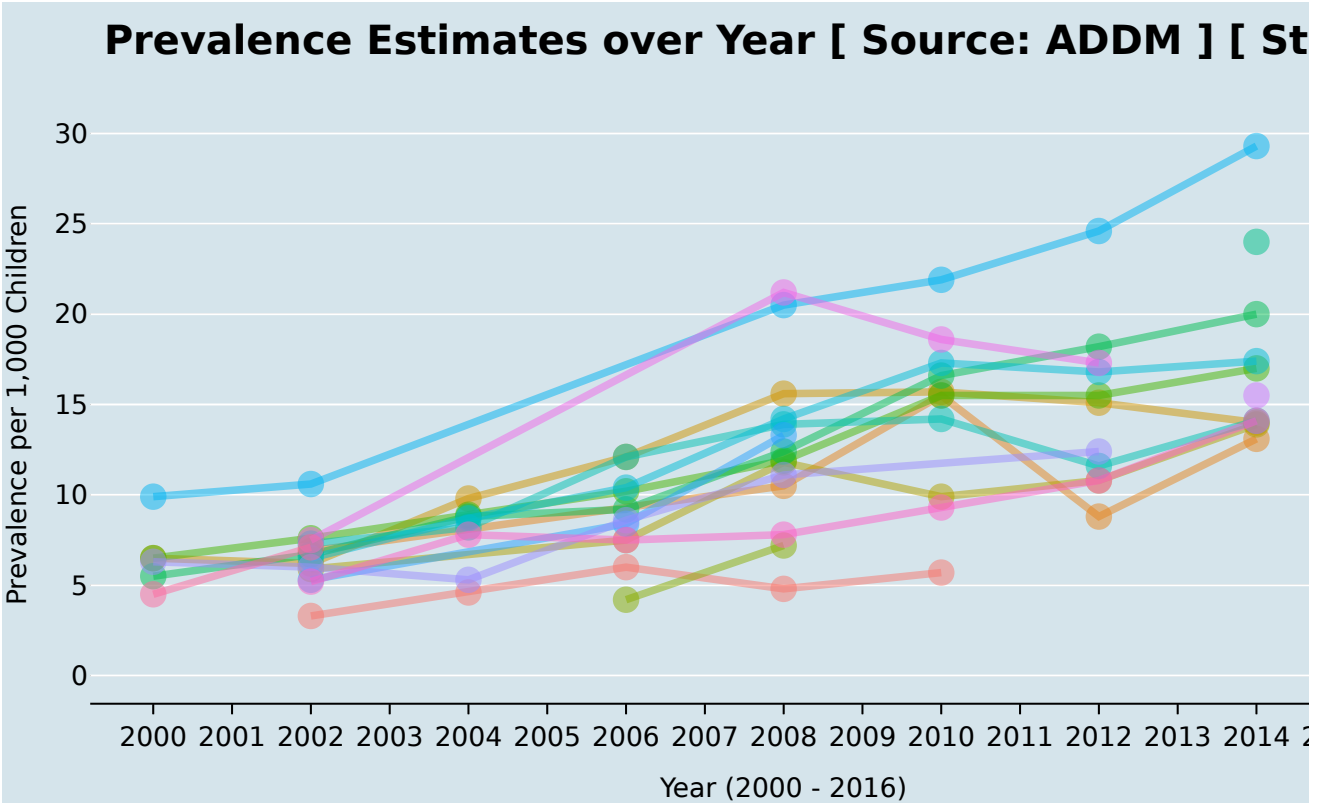
In [135]:
```
p <- ggplot(ASD_State_ADDM, aes(x = Year, y = Prevalence))
# Select (add) line chart type:
p <- p + geom_line(aes(color = State_Full2),
                   linetype = "solid",  # http://sape.inf.usi.ch/quick-referen
                   size=1,
                   alpha=0.5)
# Select (add) points to chart:
p <- p + geom_point(aes(color = State_Full2),
                    size=3,
                    shape=20,
                    alpha=0.5)
# Show plot
# p
# Customize line color and legend name:
p <- p + labs(color = "US. State")
# Adjust x and y axis, scale, limit and labels:
p <- p + scale_y_continuous(name = "Prevalence per 1,000 Children",
                            breaks = seq(0, 30, 5),
                            limits=c(0, 30)) +
  scale_x_continuous(name = "Year (2000 - 2016)",
                     breaks = seq(2000, 2016, 1),
                     limits = c(2000, 2016))
# Customize chart title:
p <- p + ggtitle("Prevalence Estimates over Year [ Source: ADDM ] [ State: ALL
# Customize chart title and axis labels:
p <- p + theme(title = element_text(face = 'bold.italic', color = "darkslategr
               axis.title = element_text(face = 'plain', color = "darkslategre
               legend.position="right")
```

In [136]:
```
# Show plot
p
```



In [137]:
```
# Dynamic chart
p_dynamic <- p + theme_economist() + scale_colour_economist() + scale_colour_d
p_dynamic <- ggplotly(p_dynamic)
p_dynamic
```

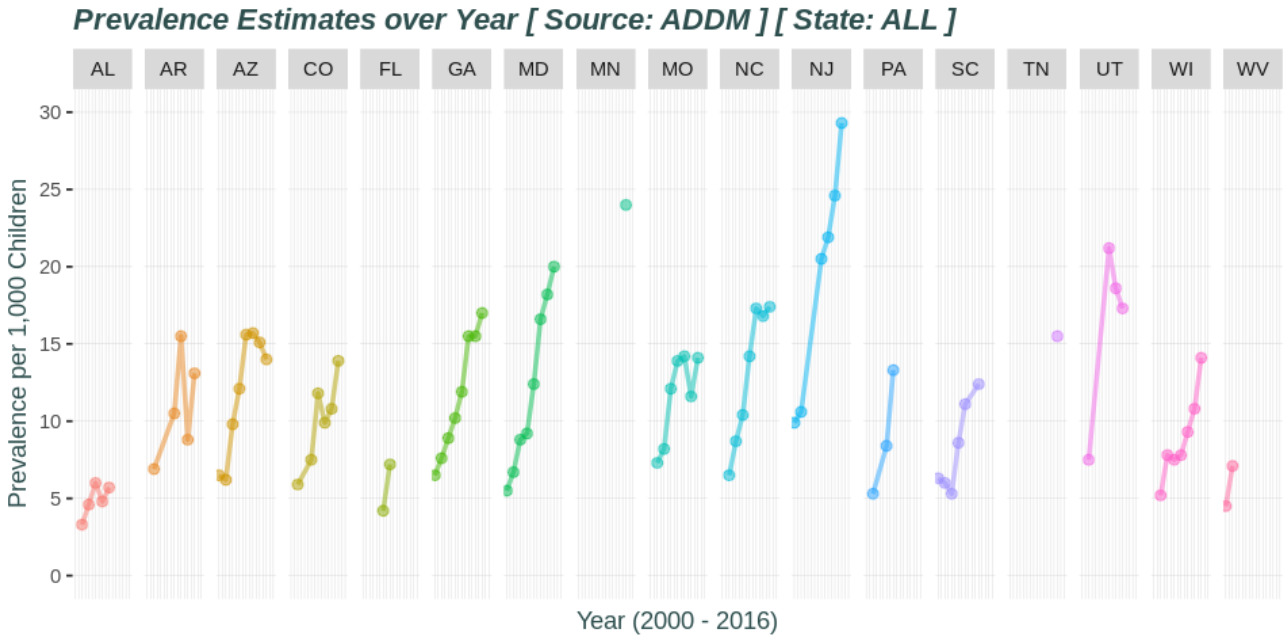
```
Scale for 'colour' is already present. Adding another scale for 'colour',
which will replace the existing scale.
```



Split chart by state

In [138]:
```
# Show plot in facet_grid
p + facet_grid(facets = . ~ State) +
  theme(legend.position = "none", # Hide legend
        axis.text.x=element_blank(),  # Hide axis
        axis.ticks.x=element_blank(), # Hide axis
        panel.background = element_blank(), # Remove panel background
        panel.grid.major = element_line(size = 0.1, linetype = 1, colour = "li
  )
```

```
geom_path: Each group consists of only one observation. Do you need to adjust
the group aesthetic?
geom_path: Each group consists of only one observation. Do you need to adjust
the group aesthetic?
```

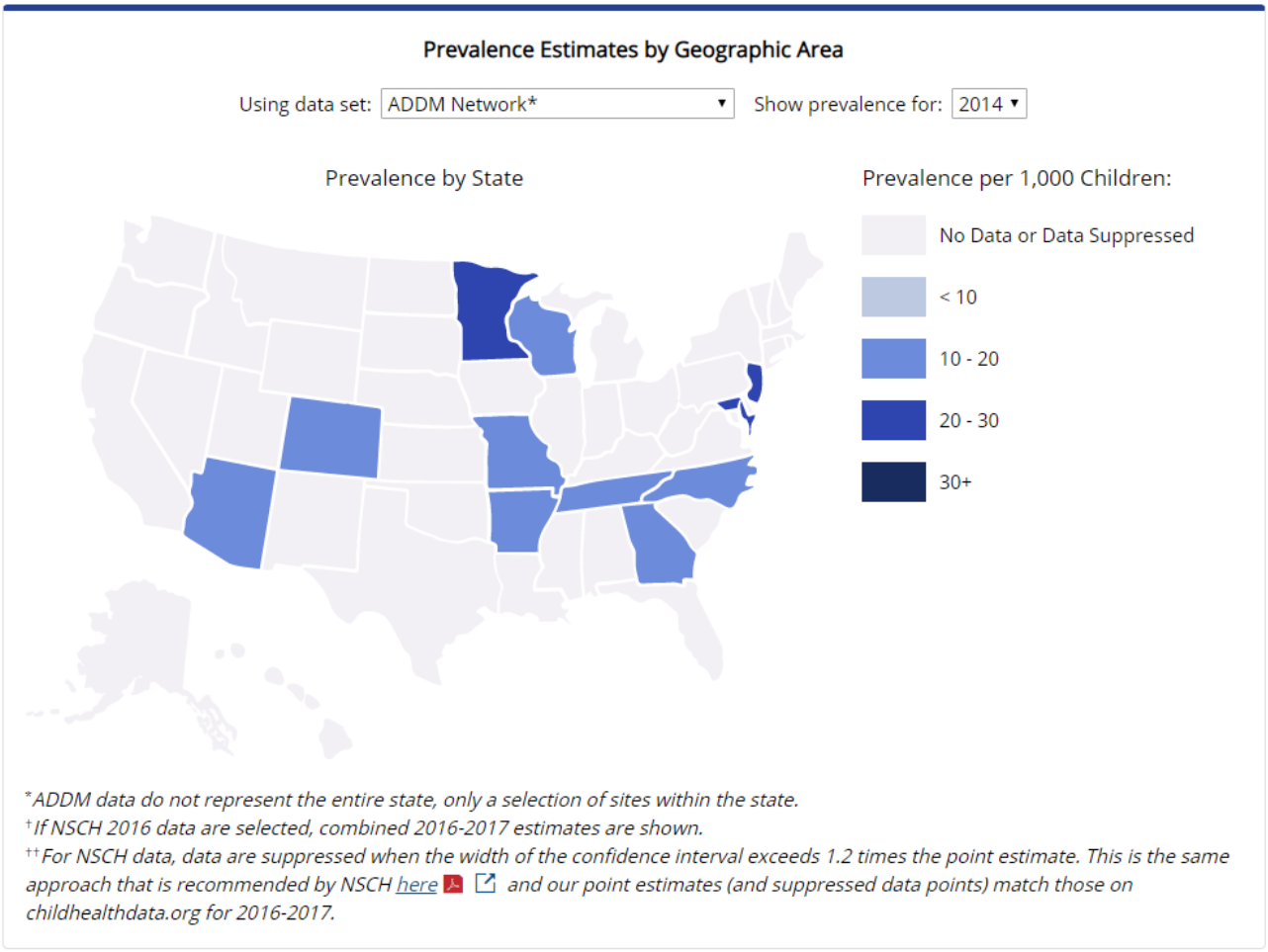


Data Visualisation (Enhanced) - Plotting on Map

In [139]:
```
# ---------------------------------
# EDA - Visualisation on map
# ---------------------------------
if(!require(usmap)){install.packages("usmap")}
library(usmap) # usmap: Mapping the US
```

```
Loading required package: usmap
```

Data Visualisation (Enhanced) - Plotting on Map [CDC] REPORTED PREVALENCE VARIES BY GEOGRAPHIC LOCATION

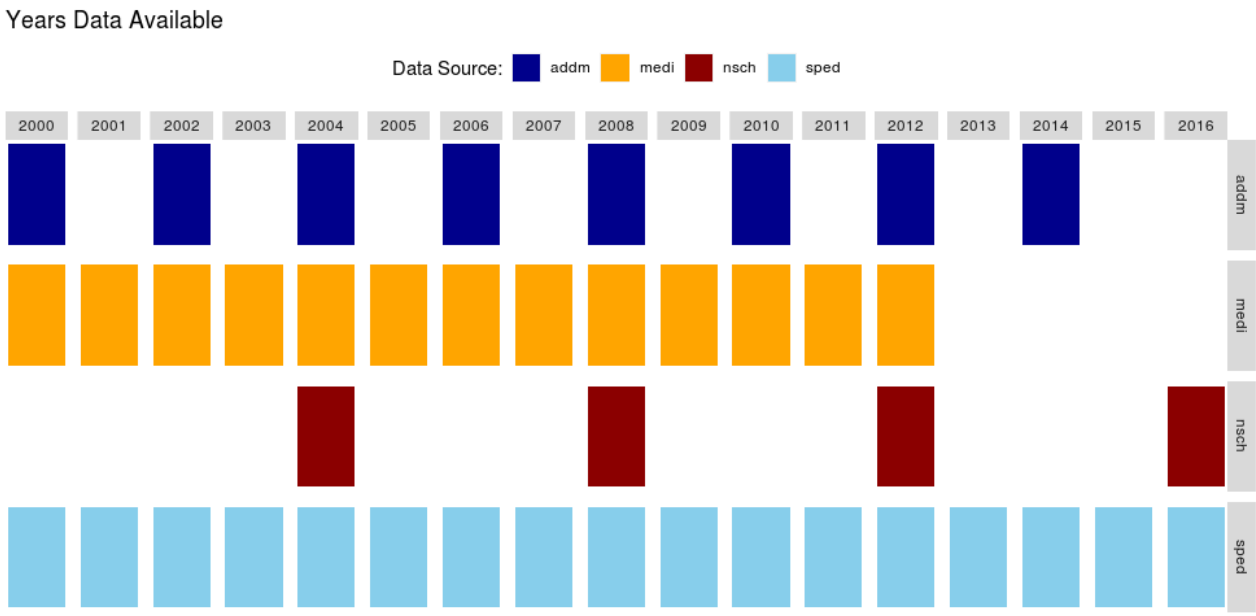


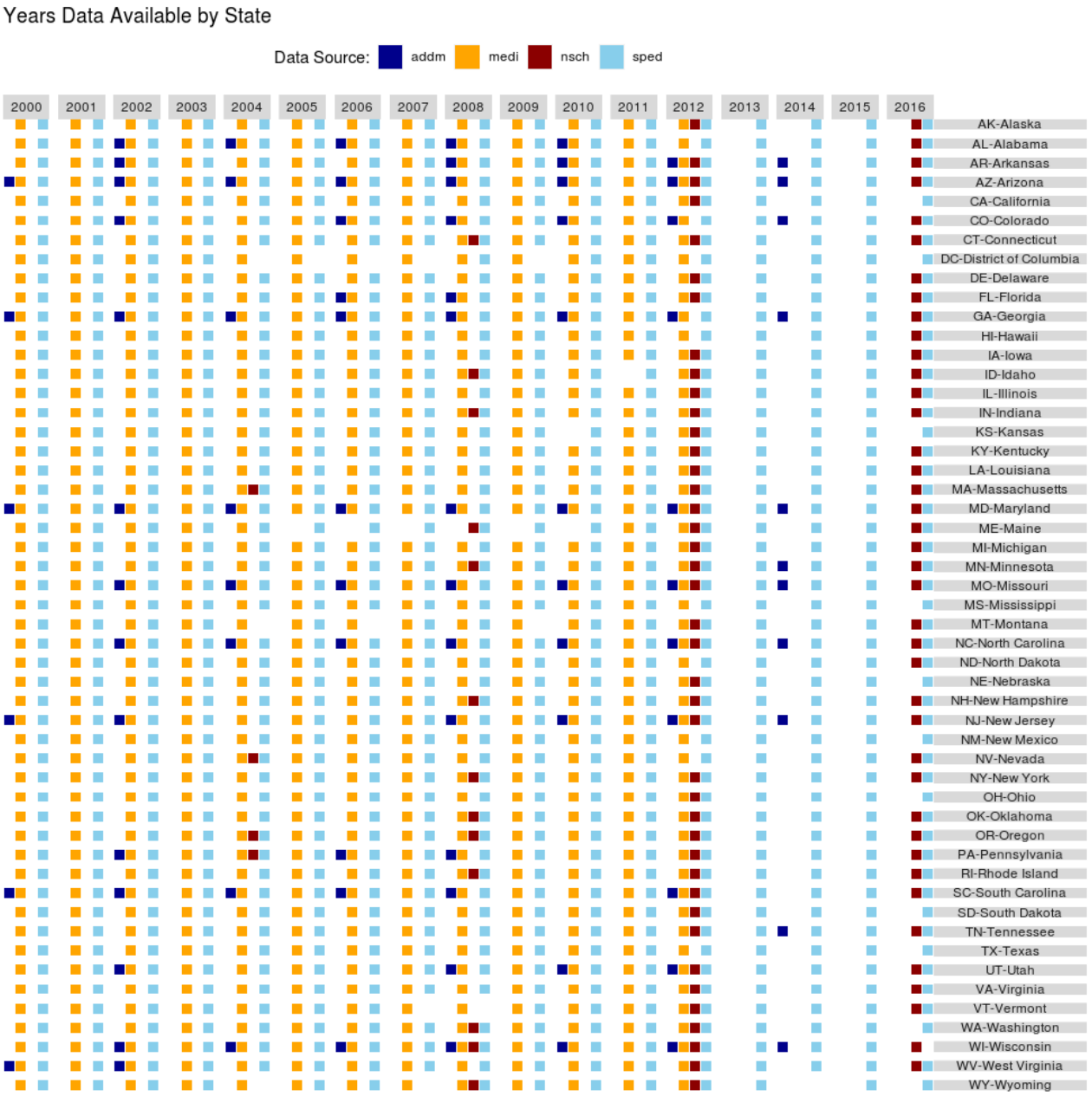
*ADDM data do not represent the entire state, only a selection of sites within the state.
†If NSCH 2016 data are selected, combined 2016-2017 estimates are shown.
††For NSCH data, data are suppressed when the width of the confidence interval exceeds 1.2 times the point estimate. This is the same approach that is recommended by NSCH here and our point estimates (and suppressed data points) match those on childhealthdata.org for 2016-2017.

Data Visualisation (Enhanced) - Plotting on Map [R] REPORTED PREVALENCE VARIES BY GEOGRAPHIC LOCATION

Let's review data availability by data Sources & Years:

- ASD_State_ADDM in Years: 2000, 2002, 2004, 2006, 2008, 2010, 2012, 2014
- ASD_State_MEDI in Years: 2000 ~ 2012
- ASD_State_NSCH in Years: 2004, 2008, 2012, 2016
- ASD_State_SPED in Years: 2000 ~ 2016





Data Visualisation (Enhanced) - Plotting on Map [R] REPORTED PREVALENCE VARIES BY GEOGRAPHIC LOCATION [Source: ADDM] [Year: 2014]

In [140]:
```
# Adjust in-line plot size to M x N
# options(repr.plot.width=8, repr.plot.height=4)
```

Prepare US State level data: [Source: ADDM] [Year: 2014]

In [141]:
```
# Prepare data - addm 2014
Map_Data_Source = 'addm' # Available values lowercase: 'addm', 'medi', 'nsch',
Map_Data_Value = 'Prevalence' # variable must be numeric, variable name in 'qu

# Uncomment below to use Prevalence of different groups:
# Map_Data_Value = 'Male.Prevalence' # variable must be numeric, variable name
# Map_Data_Value = 'Female.Prevalence' # variable must be numeric, variable na
# Map_Data_Value = 'Asian.or.Pacific.Islander.Prevalence' # variable must be n

Map_Data_Year = 2014 # must be integer
ASD_State_Subset = subset(ASD_State, Source == Map_Data_Source & Year == Map_D
```

The usmap package/function requires input data to have a column of state, or fips. (case sensitive)

- state: Name of US state
- fips: FIPS code for either a US state

https://cran.r-project.org/web/packages/usmap/vignettes/mapping.html (https://cran.r-project.org/web/packages/usmap/vignettes/mapping.html)

https://cran.r-project.org/web/packages/usmap/usmap.pdf (https://cran.r-project.org/web/packages/usmap/usmap.pdf)

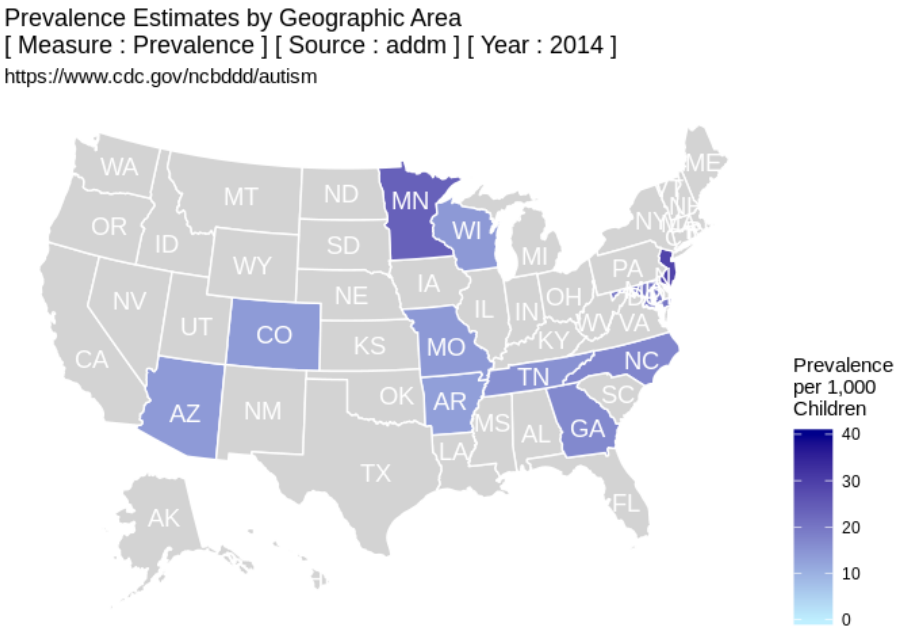
In [142]:
```
# The usmap package/function requires input data to have a column of 'state',
ASD_State_Subset$state = ASD_State_Subset$State
# Glance
head(ASD_State_Subset)
```

State	Denominator	Prevalence	Lower.CI	Upper.CI	Year	Source	Source_Full1	State_Full1	State_Full
AZ	24952	14.0	12.6	15.5	2014	addm	Autism & Developmental Disabilities Monitoring Network	Arizona	AZ-Arizona
AR	39992	13.1	12.0	14.2	2014	addm	Autism & Developmental Disabilities Monitoring Network	Arkansas	AR Arkansa
CO	41128	13.9	12.8	15.1	2014	addm	Autism & Developmental Disabilities Monitoring Network	Colorado	CO Colorad
GA	51161	17.0	15.9	18.1	2014	addm	Autism & Developmental Disabilities Monitoring Network	Georgia	GA-Georgi
MD	9955	20.0	17.4	22.9	2014	addm	Autism & Developmental Disabilities Monitoring Network	Maryland	MD Marylan
MN	9767	24.0	21.1	27.2	2014	addm	Autism & Developmental Disabilities Monitoring Network	Minnesota	MN Minnesot

Visualise: **Prevalence Estimates by Geographic Area** [Source: ADDM] [Year: 2014]

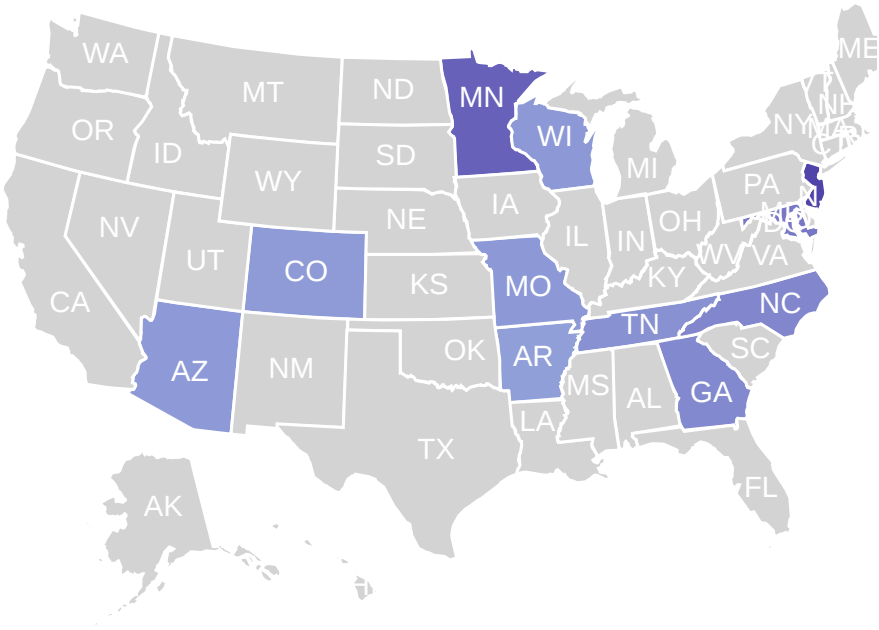
In [143]:

```
# Show data on map
p_map_addm_2014 <- plot_usmap(data = ASD_State_Subset, values = Map_Data_Value
                              color = "white", # map line colour
                              labels = TRUE,  # State name shown
                              label_color = 'white' # State name colour
) +
  scale_fill_continuous(
    na.value = "lightgrey", # Set colour with no State data
    low="lightblue1", high = "darkblue",
    name = "Prevalence\nper 1,000\nChildren",
    limits=c(0, 40) #same colour levels/limits for plots
  ) +
  labs(title = paste("Prevalence Estimates by Geographic Area", '\n[ Measure :
       subtitle = 'https://www.cdc.gov/ncbddd/autism'
  ) +
  theme(panel.background = element_rect(color = "white", fill = "white"),
        legend.position = "right")
# Show map
p_map_addm_2014
```



In [144]:
```
# Dynamic map
p_dynamic <- p_map_addm_2014
p_dynamic <- ggplotly(p_dynamic)
p_dynamic
```

Prevalence Estimates by Geographic Area
[Measure : Prevalence] [Source : addm] [Year : 2014]



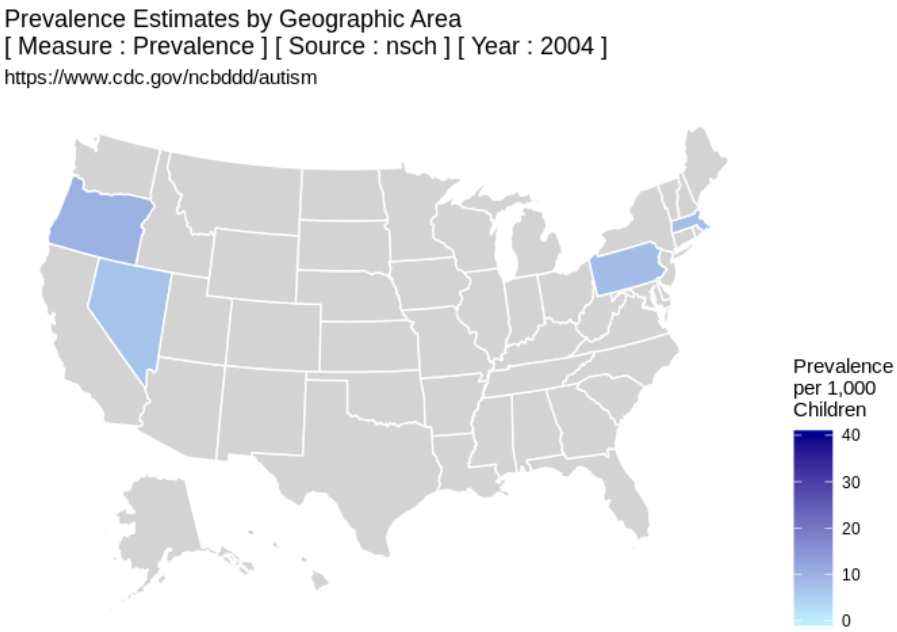
Data Visualisation (Enhanced) - Plotting on Map [R] REPORTED PREVALENCE VARIES BY GEOGRAPHIC LOCATION [Source: NSCH] [Year: 2004, 2008, 2012, 2016]

Prepare US State level data: [Source: NSCH] [Year: ALL]

In [145]:
```
Map_Data_Source = 'nsch' # Available values lowercase: 'addm', 'medi', 'nsch',
Map_Data_Value = 'Prevalence' # variable must be numeric, variable name in 'qu
```

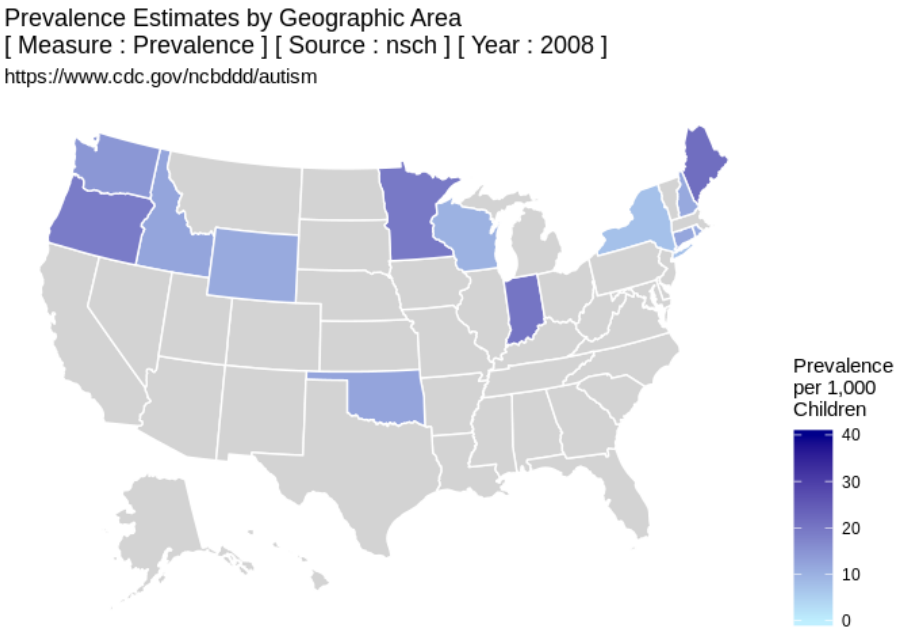
Visualise: **Prevalence Estimates by Geographic Area** [Source: NSCH] [Year: 2004]

In [146]:
```
# Prepare data - nsch 2004
Map_Data_Year = 2004 # must be integer
ASD_State_Subset = subset(ASD_State, Source == Map_Data_Source & Year == Map_D
ASD_State_Subset$state = ASD_State_Subset$State
# Plot on map
p_map_nsch_2004 <- plot_usmap(data = ASD_State_Subset, values = Map_Data_Value
p_map_nsch_2004
```



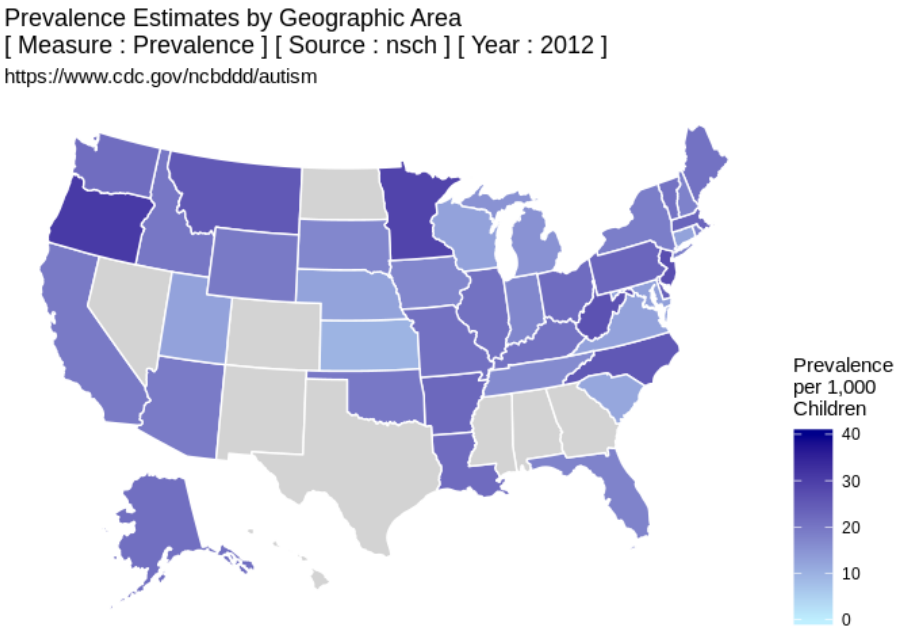
Visualise: **Prevalence Estimates by Geographic Area** [Source: NSCH] [Year: 2008]

In [147]:
```
# Prepare data - nsch 2008
Map_Data_Year = 2008 # must be integer
ASD_State_Subset = subset(ASD_State, Source == Map_Data_Source & Year == Map_D
ASD_State_Subset$state = ASD_State_Subset$State
p_map_nsch_2008 <- plot_usmap(data = ASD_State_Subset, values = Map_Data_Value
p_map_nsch_2008
```



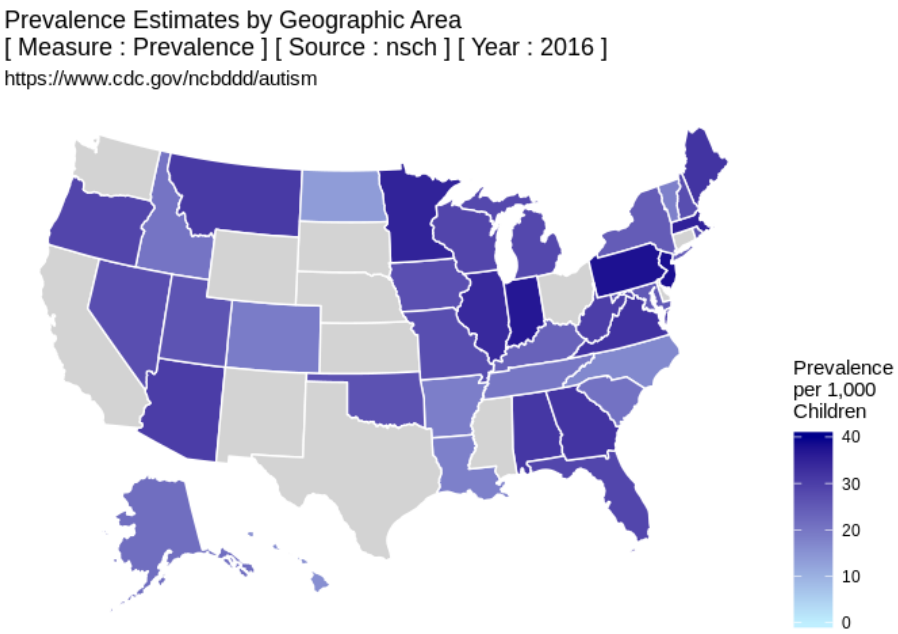
Visualise: **Prevalence Estimates by Geographic Area** [Source: NSCH] [Year: 2012]

In [148]:
```
# Prepare data - nsch 2012
Map_Data_Year = 2012 # must be integer
ASD_State_Subset = subset(ASD_State, Source == Map_Data_Source & Year == Map_D
ASD_State_Subset$state = ASD_State_Subset$State
p_map_nsch_2012 <- plot_usmap(data = ASD_State_Subset, values = Map_Data_Value
p_map_nsch_2012
```



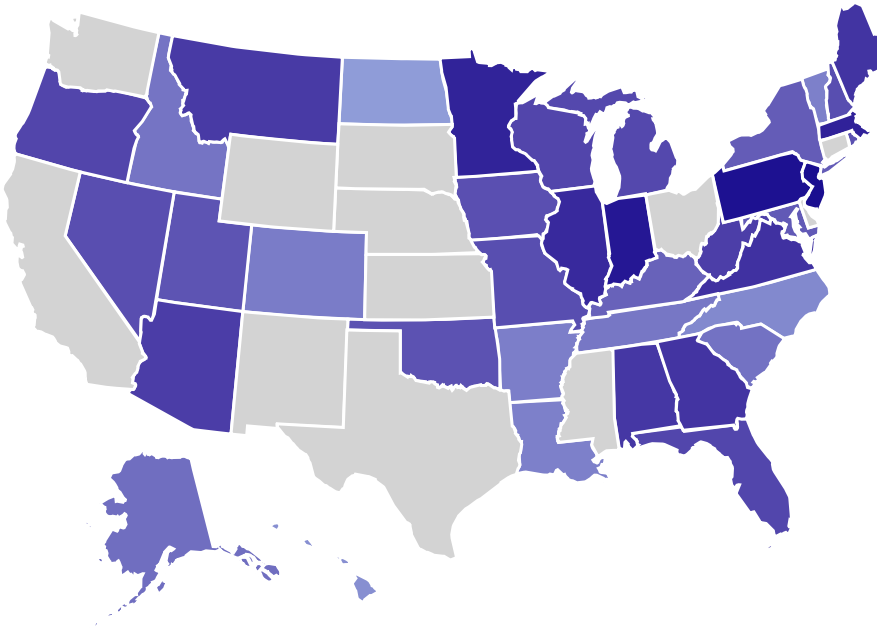
Visualise: **Prevalence Estimates by Geographic Area** [Source: NSCH] [Year: 2016]

In [149]:
```
# Prepare data - nsch 2016
Map_Data_Year = 2016 # must be integer
ASD_State_Subset = subset(ASD_State, Source == Map_Data_Source & Year == Map_D
ASD_State_Subset$state = ASD_State_Subset$State
p_map_nsch_2016 <- plot_usmap(data = ASD_State_Subset, values = Map_Data_Value
p_map_nsch_2016
```



In [150]:
```
# Dynamic map
p_dynamic <- p_map_nsch_2016 # [ Source: NSCH ] [ Year: 2016 ]
p_dynamic <- ggplotly(p_dynamic)
p_dynamic
```

Prevalence Estimates by Geographic Area
[Measure : Prevalence] [Source : nsch] [Year : 2016]

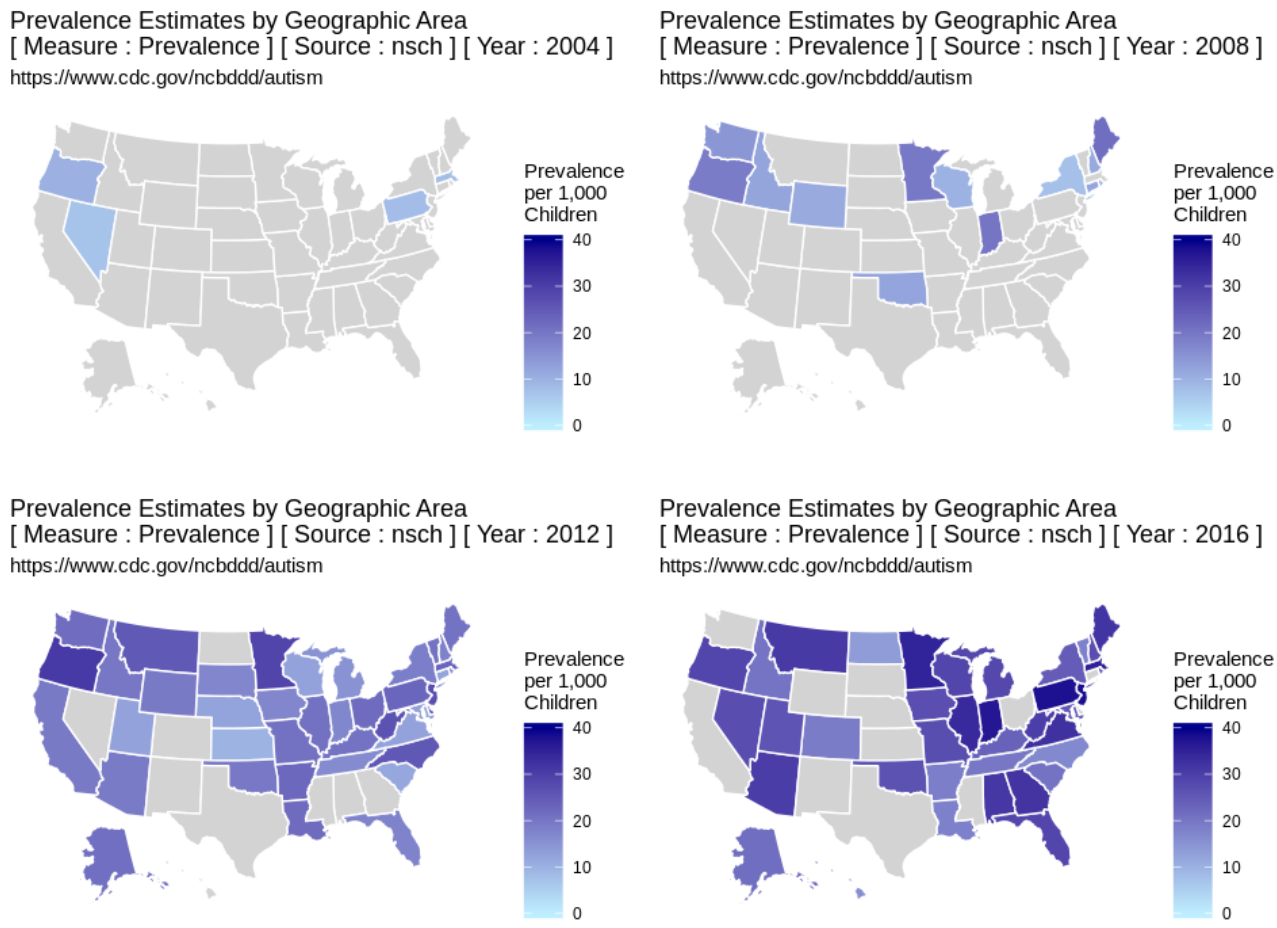


Combine multiple plots to show in one page/screen:

In [151]:
```
# Adjust in-line plot size to M x N
options(repr.plot.width=8, repr.plot.height=6)
```

In [152]:
```
# ---------------------------------
# Combine multiple plots
# ---------------------------------
if(!require(cowplot)){install.packages("cowplot")}
library('cowplot')
cowplot::plot_grid(
  p_map_nsch_2004,
  p_map_nsch_2008,
  p_map_nsch_2012,
  p_map_nsch_2016,
  nrow = 2)
```

```
Loading required package: cowplot

************************************************************
Note: As of version 1.0.0, cowplot does not change the
  default ggplot2 theme anymore. To recover the previous
  behavior, execute:
  theme_set(theme_cowplot())
************************************************************


Attaching package: ‘cowplot’

The following object is masked from ‘package:ggthemes’:

    theme_map
```



Export current plot as image file:

In [153]:
```
# ---------------------------------
# Export current plot as image file
# ---------------------------------
ggsave("plot Map Prevalence Estimates by Geographic Area [NSCH] [2004-2016].pn
       width = 60, height = 30, units = 'cm')
```

()

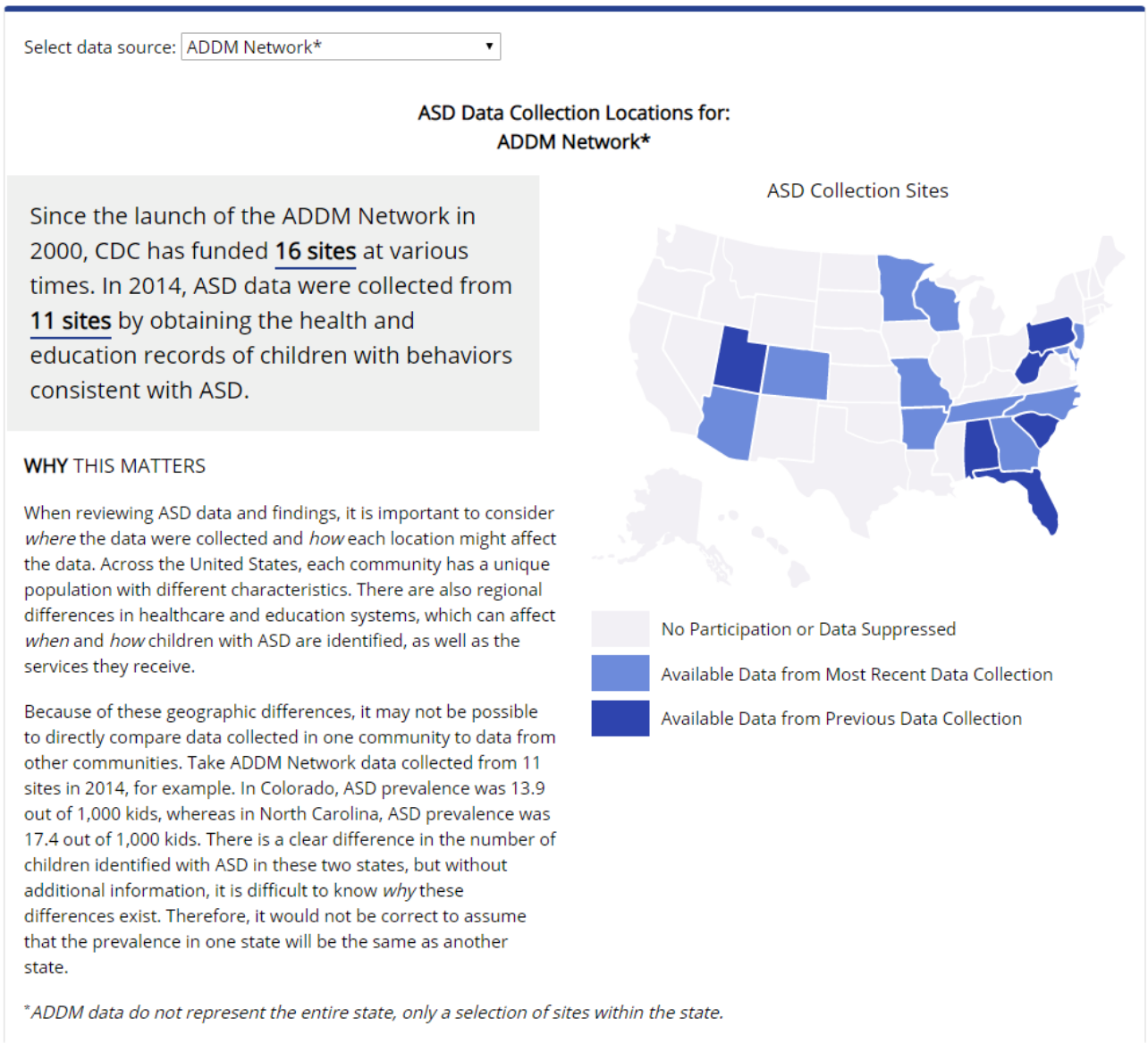
Workshop Submission

What to submit?

Choose one of below visualisations/charts, use R to construct the chart nicely.

Optionally, enhance it with additional data dimensions to be better than original chart.

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



2014 ADDM NETWORK DATA

In this section, explore the most recent ADDM data, both overall and among certain demographic groups by study area.

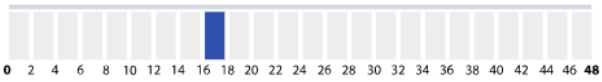
Select a location: U.S. or Total† ▾

MOST RECENT STUDY YEAR: 2014

ASD PREVALENCE PER 1,000 8-YEAR-OLD CHILDREN

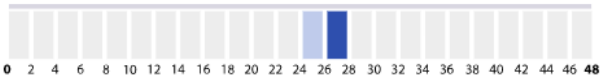
Prevalence **Overall**

Overall: 16.8 | **Lower CI**: 16.4 | **Upper CI**: 17.3

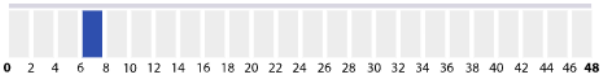


Prevalence **By Sex**

Boys: 26.6 | **Lower CI**: 25.8 | **Upper CI**: 27.4

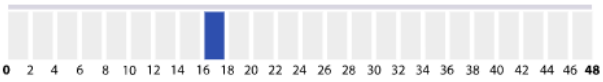


Girls: 6.6 | **Lower CI**: 6.2 | **Upper CI**: 7

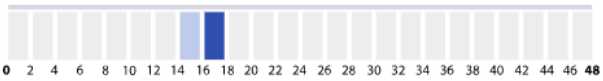


Prevalence **By Race/Ethnicity**

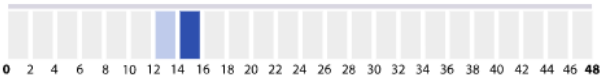
Non-Hispanic White: 17.2 | **Lower CI**: 16.5 | **Upper CI**: 17.8



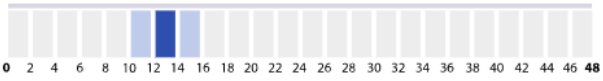
Non-Hispanic Black: 16 | **Lower CI**: 15.1 | **Upper CI**: 16.9



Hispanic: 14 | **Lower CI**: 13.1 | **Upper CI**: 14.9



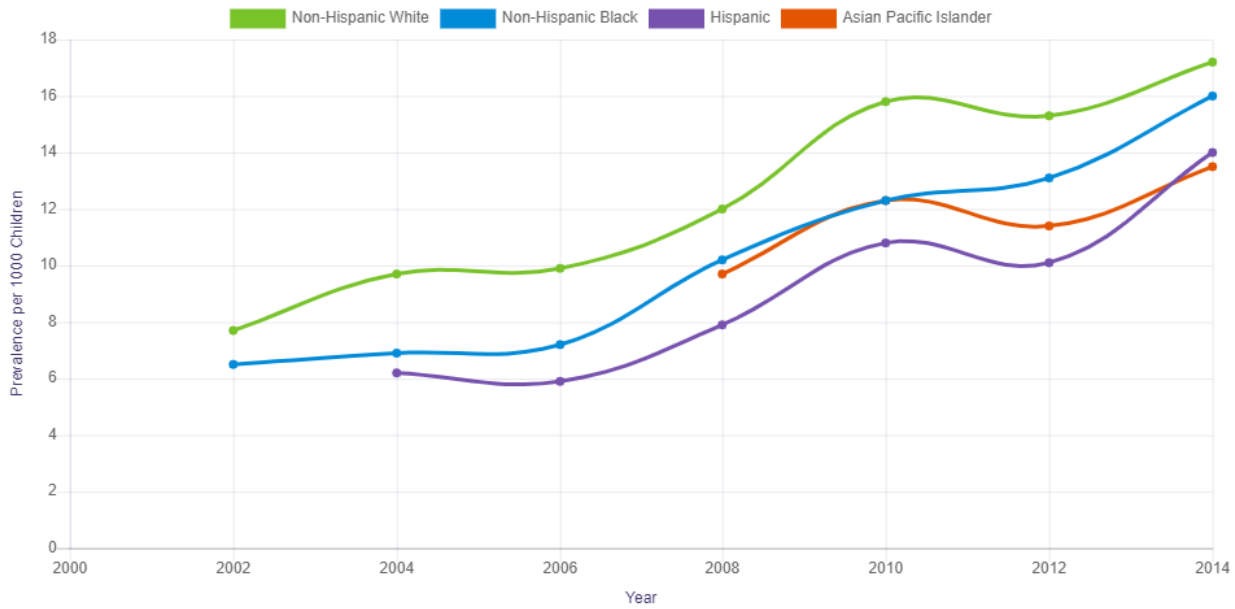
Asian/Pacific Islander: 13.5 | **Lower CI**: 11.8 | **Upper CI**: 15.4



†*ADDM estimate = the total for all sites combined.*

Prevalence Estimates by Race/Ethnicity

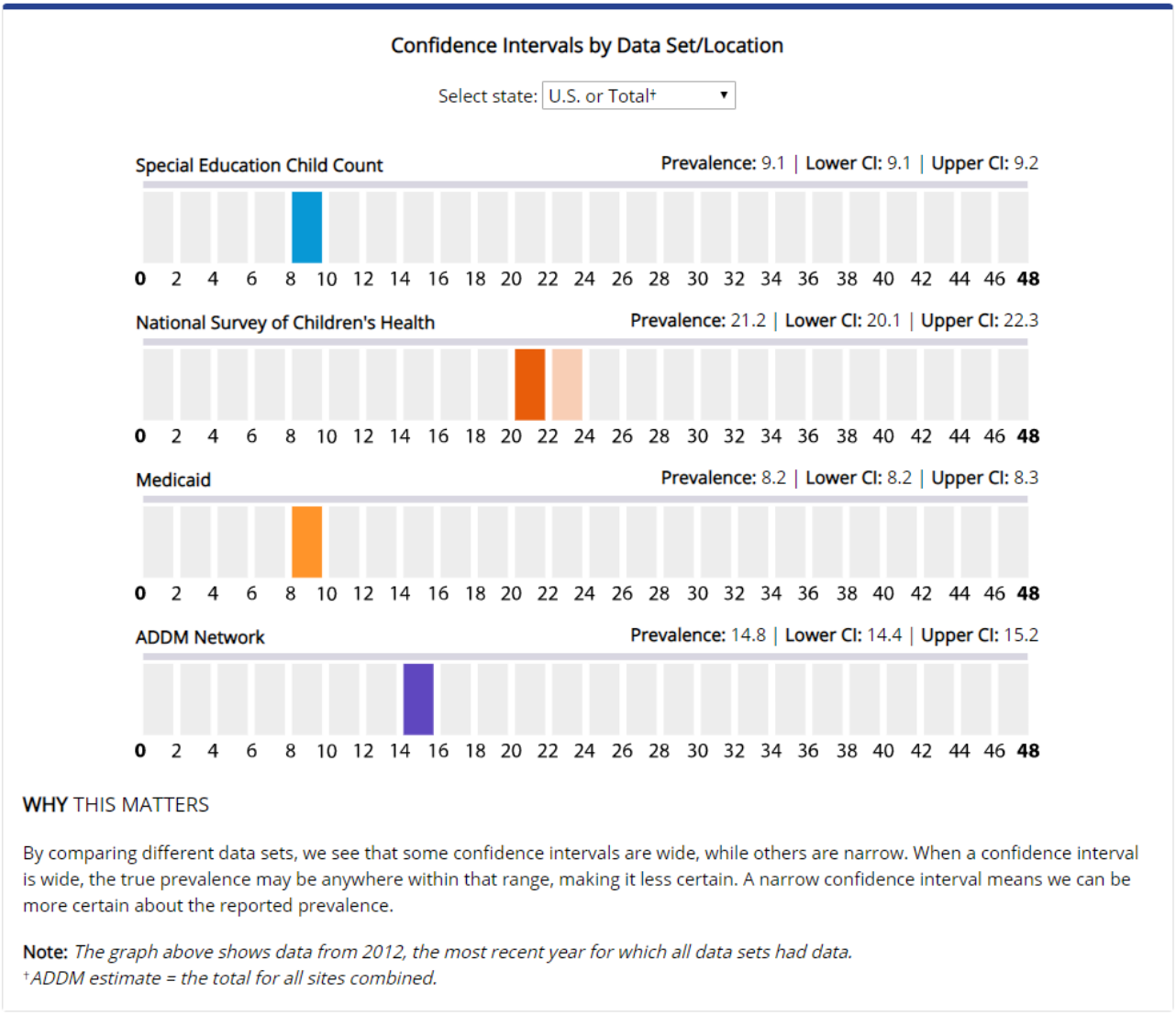
Show ADDM prevalence estimates* by race/ethnicity for: U.S. or Total† ▾



Note: *Click the icons and racial/ethnic groups above the chart to hide or unhide data. Hover your mouse over data points to show prevalence by year.*

**ADDM data do not represent the entire state, only a selection of sites within the state.*

†*ADDM estimate = the total for all sites combined.*



In [154]:
```
# Write your code below and press Shift+Enter to execute
```

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.

of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

()

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)

Correlation of Numeric Variables

In [155]:
```
# ---------------------------------
# Correlation of Numeric Variables
# ---------------------------------
cor_df = select_if(ASD_State, is.numeric) # Select only numeric variables
cor_df = cor_df[, colSums(is.na(cor_df)) == 0] #  Select vaariables without NA

# Compute correlation matrix for No-NA numeric variables:
cor_table = cor(cor_df)
cor_table
```

	Denominator	Prevalence	Lower.CI	Upper.CI	Year	Numerato
Denominator	1.00000000	-0.1374662	-0.07863304	-0.17389486	0.02851671	0.82
Prevalence	-0.13746621	1.0000000	0.95813468	0.96568034	0.64002950	0.11
Lower.CI	-0.07863304	0.9581347	1.00000000	0.85132455	0.67690938	0.21
Upper.CI	-0.17389486	0.9656803	0.85132455	1.00000000	0.56480277	0.02
Year	0.02851671	0.6400295	0.67690938	0.56480277	1.00000000	0.29
Numerator_ASD	0.82429404	0.1121787	0.21429644	0.02005452	0.29628163	1.00
Numerator_NonASD	0.99999025	-0.1392238	-0.08080949	-0.17516773	0.02638864	0.82
Proportion	-0.13735462	0.9999677	0.95851437	0.96524017	0.64020778	0.11
Chi_Wilson_Corrected_Lower.CI	-0.08734046	0.9761979	0.99597141	0.88837741	0.67167964	0.19
Chi_Wilson_Corrected_Upper.CI	-0.17380524	0.9798117	0.88384420	0.99561482	0.58775086	0.03

In [156]:
```
# ---------------------------------
# Visualise Correlation Matrix
# ---------------------------------

if(!require(corrplot)){install.packages("corrplot")}
library('corrplot')
```

```
Loading required package: corrplot
corrplot 0.84 loaded
```

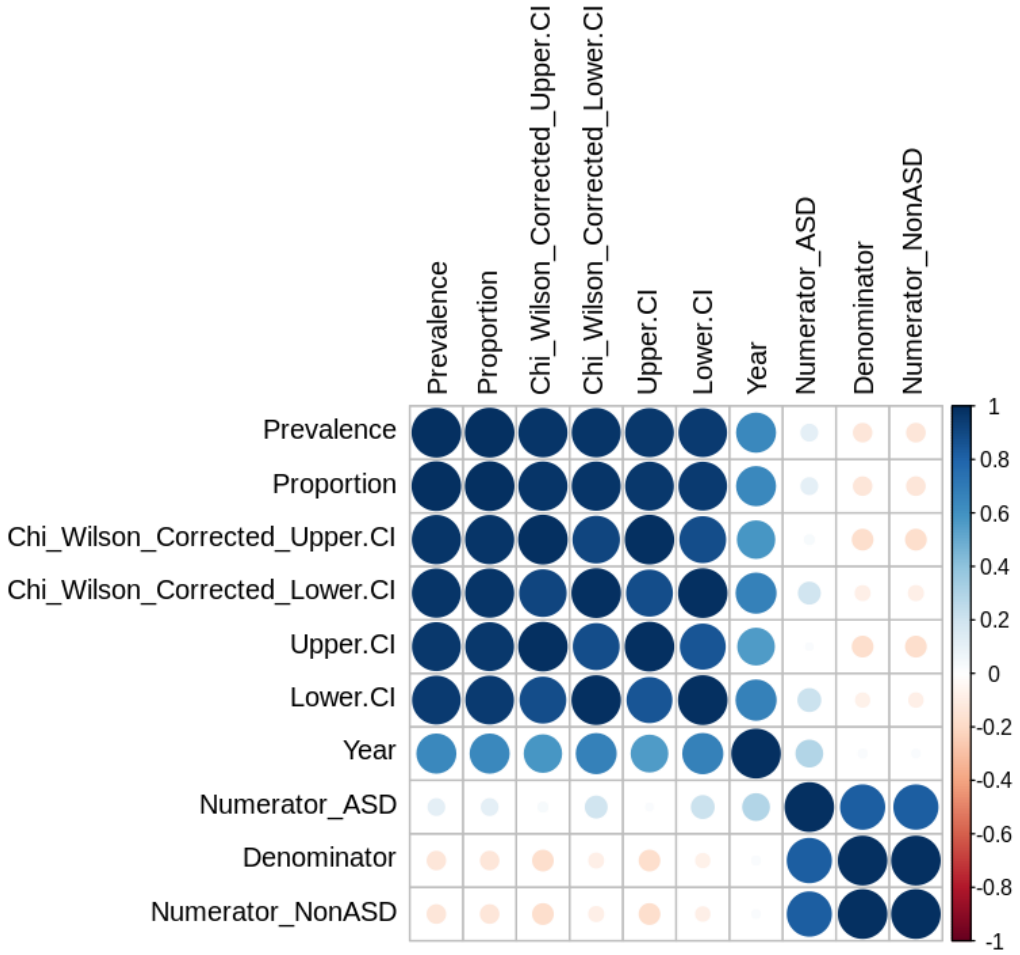
In [157]:
```
# Sort on decreasing correlations with Prevalence
cor_table_sorted <- as.matrix(sort(cor_table[,'Prevalence'], decreasing = TRUE
#
cor_table_sorted
```

Prevalence	1.0000000
Proportion	0.9999677
Chi_Wilson_Corrected_Upper.CI	0.9798117
Chi_Wilson_Corrected_Lower.CI	0.9761979
Upper.CI	0.9656803
Lower.CI	0.9581347
Year	0.6400295
Numerator_ASD	0.1121787
Denominator	-0.1374662
Numerator_NonASD	-0.1392238

In [158]:
```
# Select corelations variables based on threshold:
#cor_var_high <- names(which(apply(cor_table_sorted, 1, function(x) abs(x)>0.2
cor_var_high <- names(which(apply(cor_table_sorted, 1, function(x) abs(x)>0.05
#
cor_var_high
```

'Prevalence' 'Proportion' 'Chi_Wilson_Corrected_Upper.CI' 'Chi_Wilson_Corrected_Lower.CI' 'Upper.CI' 'Lower.CI' 'Year' 'Numerator_ASD' 'Denominator' 'Numerator_NonASD'

In [159]:
```
# Visualise:
cor_table_plot <- cor_table[cor_var_high, cor_var_high]
# cor_table_plot
#
corrplot(cor_table_plot, tl.col="black", tl.pos = "lt")
```



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