EPID674 Epidemiologic Data Analysis using R

Graphing in R

Kelly Bakulski, Lauren Middleton

Last compiled on December 12, 2021

Table of Contents

# Set up options

# Install new packages

# Load packages

# Load data

# Check the file path  
here("nhanes\_dataset.rda")

## [1] "/cloud/project/nhanes\_dataset.rda"

# Load the saved R data  
load(here("nhanes\_dataset.rda"))

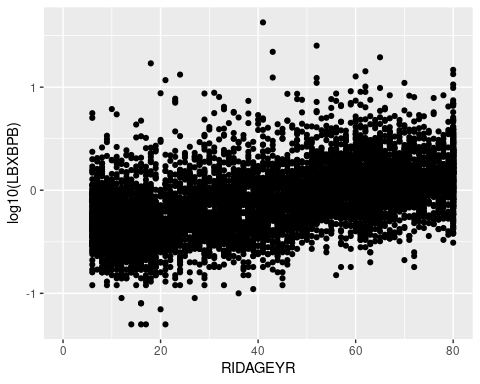
# Useful websites for ggplot information:

* Online introduction book: [**https://ggplot2-book.org/index.html**](https://ggplot2-book.org/index.html)
* For theory: [**https://rkabacoff.github.io/datavis/**](https://rkabacoff.github.io/datavis/)
* How to code: [**http://www.sthda.com/english/wiki/ggplot2-essentials**](http://www.sthda.com/english/wiki/ggplot2-essentials)

# Introduction to ggplot

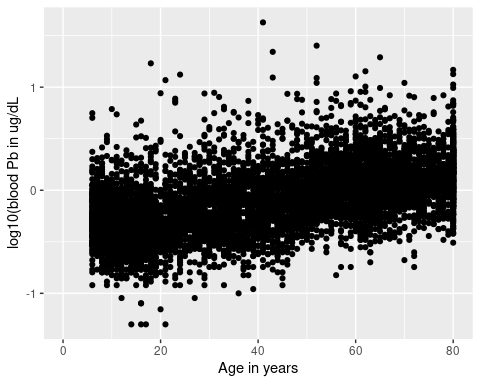
# Specify the dataset, variables on the x-axis, y-axis  
ggplot(nhanes\_dataset,  
 aes(x = RIDAGEYR,  
 y = log10(LBXBPB))) +  
 geom\_point()

## Warning: Removed 2370 rows containing missing values (geom\_point).



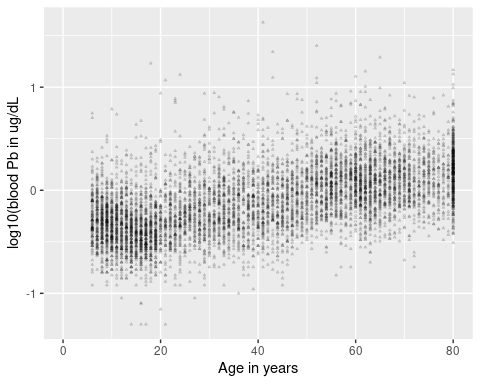
# Update axes labels  
ggplot(nhanes\_dataset,  
 aes(x = RIDAGEYR,  
 y = log10(LBXBPB))) +  
 geom\_point() +   
 labs(x = "Age in years",   
 y = "log10(blood Pb in ug/dL")

## Warning: Removed 2370 rows containing missing values (geom\_point).



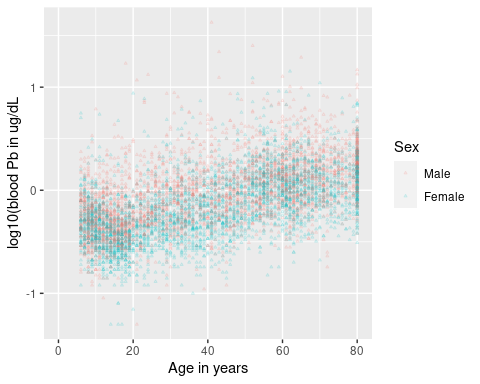
# Change point size, transparency, shape   
ggplot(nhanes\_dataset,  
 aes(x = RIDAGEYR,  
 y = log10(LBXBPB))) +  
 geom\_point(size = 0.3,   
 alpha = 0.2,   
 shape = 2) +   
 labs(x = "Age in years",   
 y = "log10(blood Pb in ug/dL")

## Warning: Removed 2370 rows containing missing values (geom\_point).



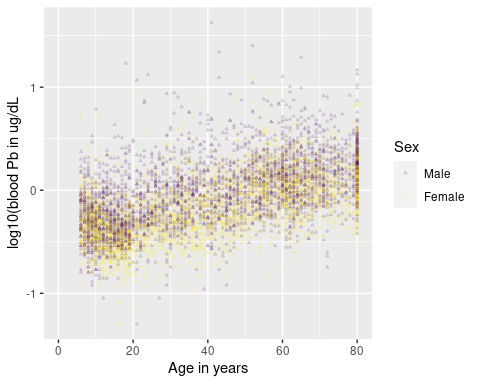
# Add color variable  
ggplot(nhanes\_dataset,  
 aes(x = RIDAGEYR,  
 y = log10(LBXBPB),  
 color = sex)) +  
 geom\_point(size = 0.4,   
 alpha = 0.2,  
 shape = 2) +   
 labs(x = "Age in years",   
 y = "log10(blood Pb in ug/dL",  
 color = "Sex")

## Warning: Removed 2370 rows containing missing values (geom\_point).



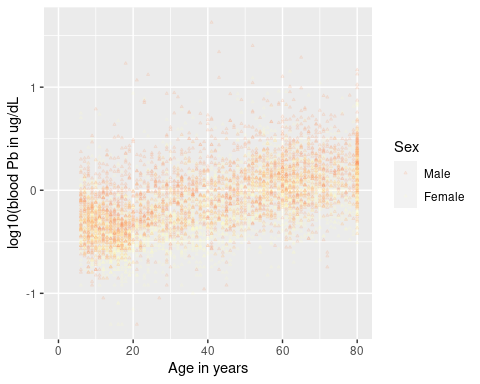
# Change color palette  
ggplot(nhanes\_dataset,  
 aes(x = RIDAGEYR,  
 y = log10(LBXBPB),  
 color = sex)) +  
 geom\_point(size = 0.4,   
 alpha = 0.2,  
 shape = 2) +   
 labs(x = "Age in years",   
 y = "log10(blood Pb in ug/dL",  
 color = "Sex") +  
 scale\_color\_viridis\_d()

## Warning: Removed 2370 rows containing missing values (geom\_point).



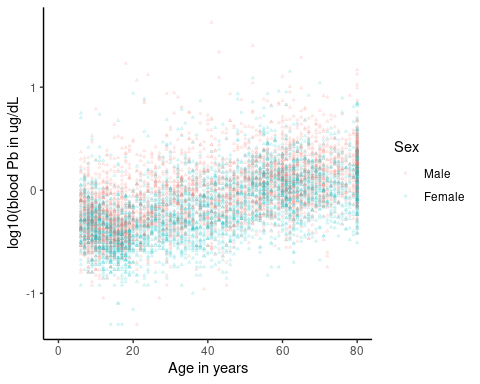
# Try another color palette  
ggplot(nhanes\_dataset,  
 aes(x = RIDAGEYR,  
 y = log10(LBXBPB),  
 color = sex)) +  
 geom\_point(size = 0.4,   
 alpha = 0.2,  
 shape = 2) +   
 labs(x = "Age in years",   
 y = "log10(blood Pb in ug/dL",  
 color = "Sex") +  
 scale\_color\_brewer(palette = "Spectral")

## Warning: Removed 2370 rows containing missing values (geom\_point).



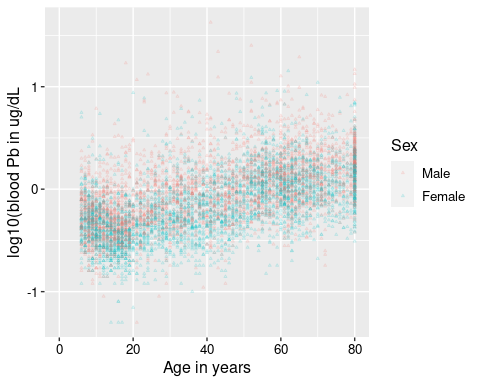
# Change theme to preset features  
ggplot(nhanes\_dataset,  
 aes(x = RIDAGEYR,  
 y = log10(LBXBPB),  
 color = sex)) +  
 geom\_point(size = 0.4,   
 alpha = 0.2,  
 shape = 2) +   
 labs(x = "Age in years",   
 y = "log10(blood Pb in ug/dL",  
 color = "Sex") +   
 theme\_classic()

## Warning: Removed 2370 rows containing missing values (geom\_point).



# Manually change plot theme features  
ggplot(nhanes\_dataset,  
 aes(x = RIDAGEYR,  
 y = log10(LBXBPB),  
 color = sex)) +  
 geom\_point(size = 0.4,   
 alpha = 0.2,  
 shape = 2) +   
 labs(x = "Age in years",   
 y = "log10(blood Pb in ug/dL",  
 color = "Sex") +   
 theme(axis.text = element\_text(color = "black", size = 10), #color and text size of axis labels  
 axis.title = element\_text(size = 12), #text size of axis titles  
 legend.title = element\_text(size = 12), #text size of legend title  
 legend.text = element\_text(size = 10)) #text size of legend labels

## Warning: Removed 2370 rows containing missing values (geom\_point).



# Check your understanding!

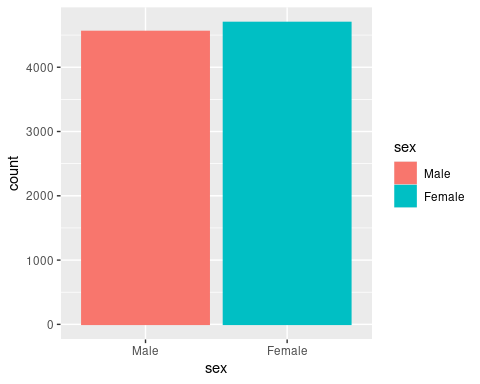
Make a scatter plot of cotinine (LBXCOT) on the x-axis and cadmium (LBXBCD) on the y-axis. Be sure to include axis labels with units. Experiment with changing the color, size, and transparency of the points.

# Make a barplot: Categorical data

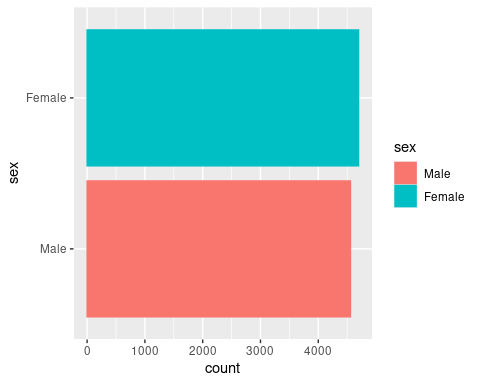
# Make a simple barplot  
ggplot(nhanes\_dataset,  
 aes(x = sex)) +  
 geom\_bar()



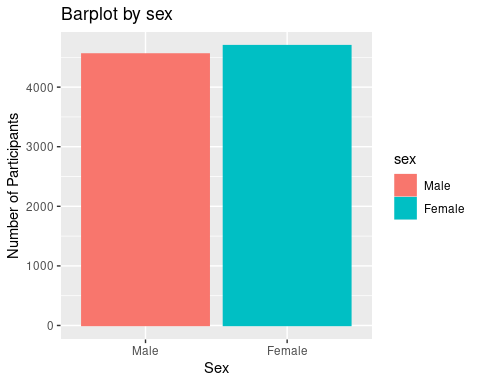
# Add some colors  
ggplot(nhanes\_dataset,  
 aes(x = sex,  
 fill = sex,  
 color = sex)) +  
 geom\_bar()



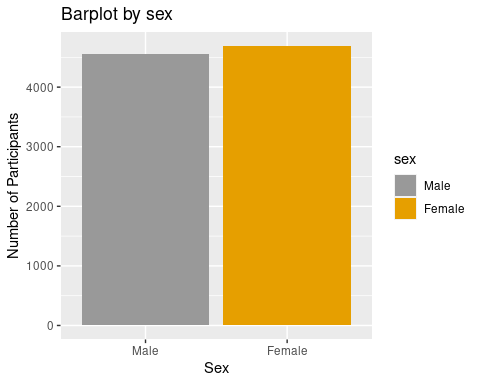
# Turn the plot horizontally  
ggplot(nhanes\_dataset,  
 aes(x = sex,  
 fill = sex,  
 color = sex)) +  
 geom\_bar() +  
 coord\_flip()



# Add labels  
ggplot(nhanes\_dataset,  
 aes(x = sex,  
 fill = sex,  
 color = sex)) +  
 geom\_bar() +  
 labs(title = "Barplot by sex",  
 x = "Sex",  
 y = "Number of Participants")



# Manually set the colors  
ggplot(nhanes\_dataset,  
 aes(x = sex,  
 fill = sex)) +  
 geom\_bar() +  
 labs(title = "Barplot by sex",  
 x = "Sex",  
 y = "Number of Participants") +  
 scale\_fill\_manual(values=c("#999999", "#E69F00"))

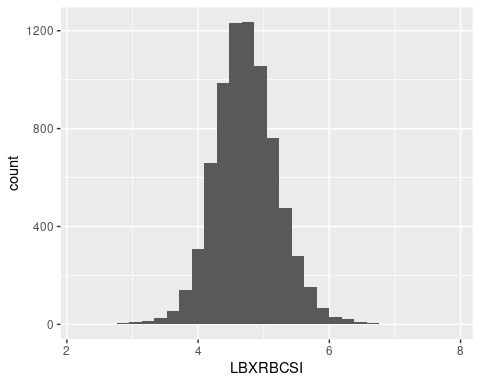


# Histogram: Numeric data

# Histogram  
ggplot(nhanes\_dataset,  
 aes(x = LBXRBCSI)) +  
 geom\_histogram()

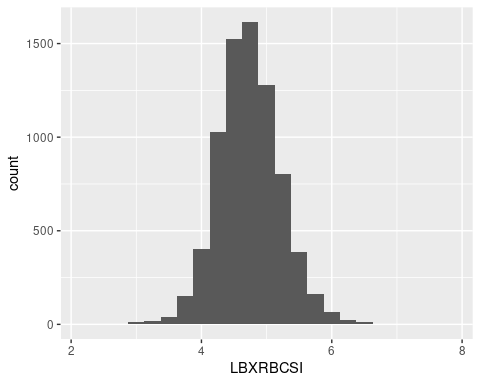
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 1726 rows containing non-finite values (stat\_bin).



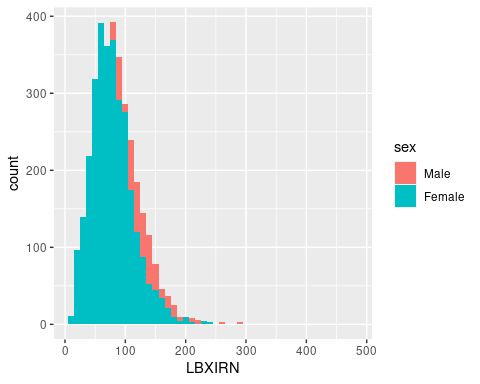
# Adjust size of boxes  
ggplot(nhanes\_dataset,  
 aes(x = LBXRBCSI)) +  
 geom\_histogram(binwidth = 0.25) +   
 labs("Red Blood Cell Count (million cells/uL)")

## Warning: Removed 1726 rows containing non-finite values (stat\_bin).



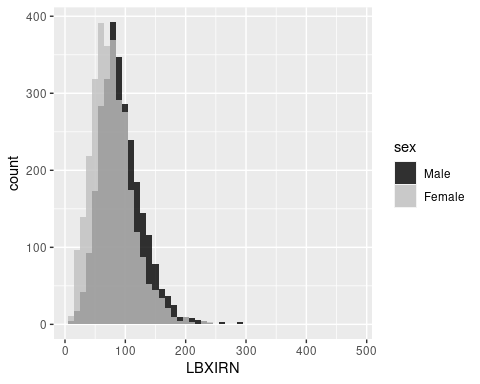
# Overlapping histograms  
ggplot(nhanes\_dataset,  
 aes(x = LBXIRN,  
 fill = sex)) + #stratify by sex  
 geom\_histogram(binwidth = 10,  
 position = "identity") #default position is stack

## Warning: Removed 3332 rows containing non-finite values (stat\_bin).



# Make a black and white plot with transparency  
ggplot(nhanes\_dataset,  
 aes(x = LBXIRN,  
 fill = sex)) +  
 geom\_histogram(binwidth = 10,  
 position = "identity",  
 alpha = 0.8) + #transparency  
 scale\_fill\_grey(start = 0,  
 end = 0.75)

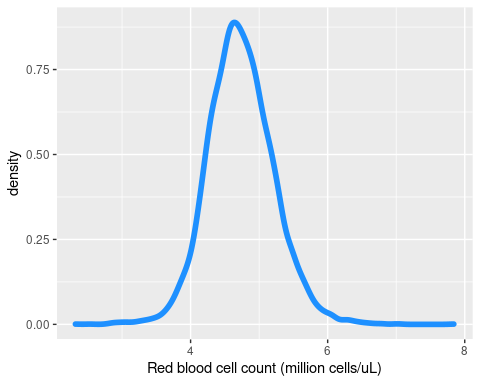
## Warning: Removed 3332 rows containing non-finite values (stat\_bin).



# Density plot

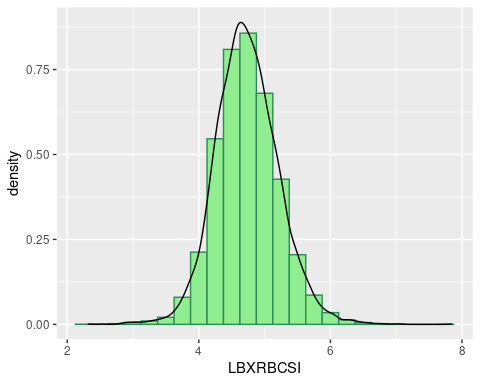
# Density plot  
ggplot(nhanes\_dataset,  
 aes(x = LBXRBCSI)) +  
 geom\_density(size=2,   
 color="dodgerblue") +  
 labs(x = "Red blood cell count (million cells/uL)")

## Warning: Removed 1726 rows containing non-finite values (stat\_density).



# Add density plot to a histogram  
ggplot(nhanes\_dataset,  
 aes(x = LBXRBCSI, y = ..density..)) +  
 geom\_histogram(fill = "lightgreen",  
 color = "seagreen",  
 binwidth = 0.25) +  
 geom\_density()

## Warning: Removed 1726 rows containing non-finite values (stat\_bin).  
  
## Warning: Removed 1726 rows containing non-finite values (stat\_density).



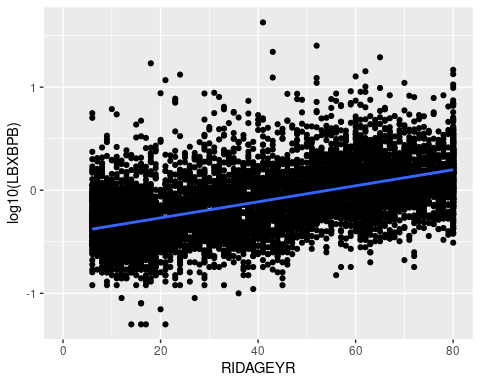
# Scatter plot with lines

# Add a linear fit  
ggplot(nhanes\_dataset,  
 aes(x = RIDAGEYR,  
 y = log10(LBXBPB))) +  
 geom\_point() +  
 geom\_smooth(method = lm)

## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 2370 rows containing non-finite values (stat\_smooth).

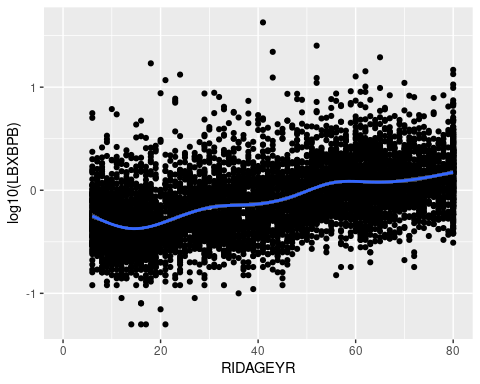
## Warning: Removed 2370 rows containing missing values (geom\_point).



# Add a smoothing line  
ggplot(nhanes\_dataset,  
 aes(x = RIDAGEYR,  
 y = log10(LBXBPB))) +  
 geom\_point() +  
 geom\_smooth()

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

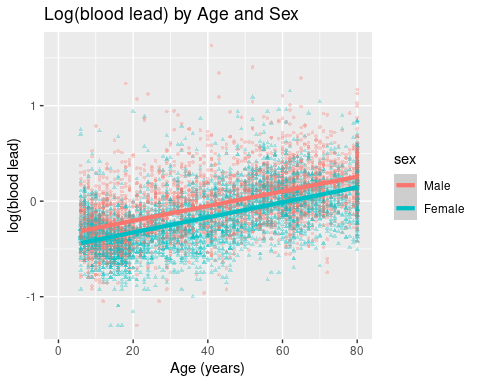
## Warning: Removed 2370 rows containing non-finite values (stat\_smooth).  
  
## Warning: Removed 2370 rows containing missing values (geom\_point).



# Split by another variable  
ggplot(nhanes\_dataset,  
 aes(x = RIDAGEYR,  
 y = log10(LBXBPB),  
 shape = sex)) +  
 geom\_point(aes(color = sex,  
 shape = sex),  
 size = 1,  
 alpha = 0.3) +  
 geom\_smooth(method = lm,  
 aes(color = sex),  
 size = 1.5) +  
 labs(title = "Log(blood lead) by Age and Sex",  
 x = "Age (years)",  
 y = "log(blood lead)")

## `geom\_smooth()` using formula 'y ~ x'

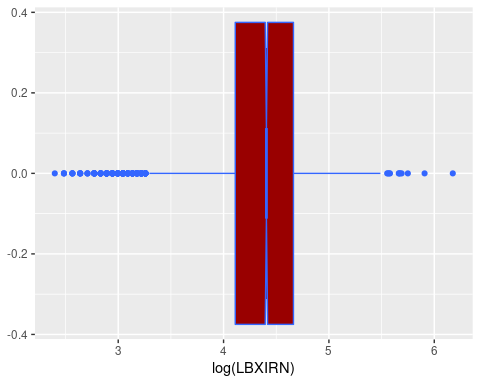
## Warning: Removed 2370 rows containing non-finite values (stat\_smooth).  
  
## Warning: Removed 2370 rows containing missing values (geom\_point).



# Boxplot

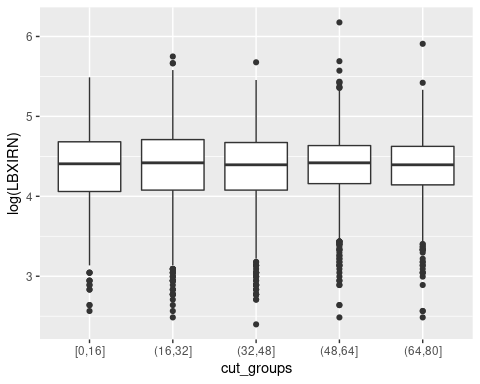
# Boxplot  
ggplot(nhanes\_dataset,  
 aes(x = log(LBXIRN))) +  
 geom\_boxplot(fill = "#990000",  
 color = "#3366FF",  
 notch = T,  
 notchwidth = .3)

## Warning: Removed 3332 rows containing non-finite values (stat\_boxplot).



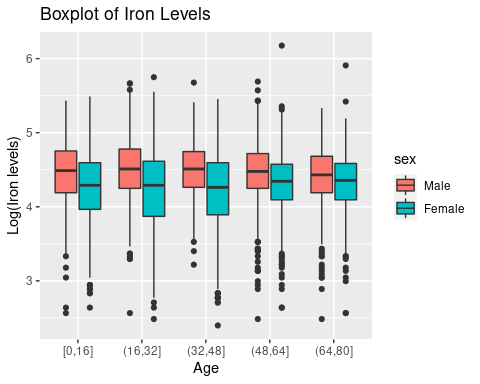
# Boxplot split by age groups  
ggplot(nhanes\_dataset,  
 aes(x = cut\_groups,  
 y = log(LBXIRN))) +  
 geom\_boxplot()

## Warning: Removed 3332 rows containing non-finite values (stat\_boxplot).



# Boxplot split by age and sex, with colors!  
ggplot(nhanes\_dataset,  
 aes(x = cut\_groups,  
 y = log(LBXIRN))) +  
 geom\_boxplot(aes(fill = sex)) +  
 labs(title = "Boxplot of Iron Levels",  
 x="Age",  
 y="Log(Iron levels)")

## Warning: Removed 3332 rows containing non-finite values (stat\_boxplot).



# Violin plot

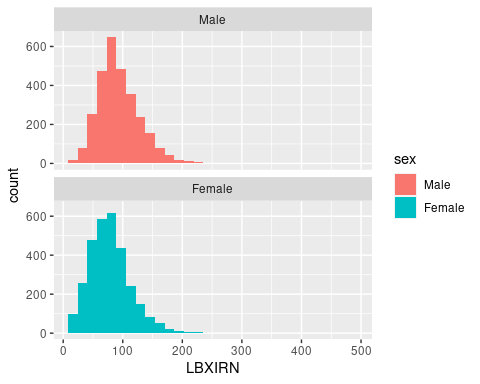
# Basic violin plot  
ggplot(nhanes\_dataset,  
 aes(x = cut\_groups,  
 y = log(LBXBPB))) +   
 geom\_violin()  
  
# Add beeswarm data points on top  
ggplot(nhanes\_dataset,  
 aes(x = cut\_groups,  
 y = log(LBXBPB))) +   
 geom\_violin() +  
 geom\_jitter(shape = 16, #circle  
 position = position\_jitter(0.2),  
 alpha = 0.25)  
  
# Add colors  
ggplot(nhanes\_dataset,  
 aes(x = cut\_groups,  
 y = log(LBXBPB),  
 fill = cut\_groups)) +   
 geom\_violin() +  
 geom\_jitter(shape = 16, #circle  
 position = position\_jitter(0.2),  
 alpha = 0.25) +  
 scale\_fill\_brewer(palette = "Blues") +  
 labs(title = "Violin plots by age group",  
 x = "Age Group (years)",  
 y = "Log(Blood Lead) (ug/dL)")

# Facet wrap: Display categories in separate plots

# Stratified histograms in separate plots  
ggplot(nhanes\_dataset,  
 aes(x = LBXIRN,  
 fill = sex)) +  
 geom\_histogram() +  
 facet\_wrap(vars(sex),  
 ncol = 1)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 3332 rows containing non-finite values (stat\_bin).



# Assign plots as objects

# Assign plots from earlier as objects  
viol\_plot <- ggplot(nhanes\_dataset,  
 aes(x = cut\_groups,  
 y = log(LBXBPB),  
 fill = cut\_groups)) +   
 geom\_violin() +  
 geom\_jitter(shape = 16, #circle  
 position = position\_jitter(0.2),  
 alpha = 0.25) +  
 scale\_fill\_brewer(palette = "Blues") +  
 labs(title = "Violin plots by age group",  
 x = "Age Group (years)",  
 y = "log(blood lead) (ug/dL)")  
  
scatter\_plot <- ggplot(nhanes\_dataset,  
 aes(x = RIDAGEYR,  
 y = log10(LBXBPB),  
 shape = sex)) +  
 geom\_point(aes(color = sex,  
 shape = sex),  
 size = 1,  
 alpha = 0.5) +  
 geom\_smooth(method = lm,  
 aes(color = sex)) +  
 labs(title = "Log(blood lead) by age and sex",  
 x = "Age (years)",  
 y = "log(blood lead)")  
  
bar\_plot <- ggplot(nhanes\_dataset,  
 aes(x = sex,  
 fill = sex)) +  
 geom\_bar() +  
 labs(title = "Barplot by sex",  
 x = "Sex",  
 y = "Number of Participants") +  
 scale\_fill\_manual(values=c("#999999", "#E69F00"))  
  
facet\_plot <- ggplot(nhanes\_dataset,  
 aes(x = LBXIRN,  
 fill = sex)) +  
 geom\_histogram() +  
 facet\_wrap(vars(sex),  
 ncol = 1) +   
 labs(title = "Histogram of iron by sex",  
 x = "Serum Iron in ug/dL")

# Combine multiple types of plots on one page

# View the four plots on one page  
ggarrange(bar\_plot, viol\_plot, scatter\_plot, facet\_plot,  
 ncol = 2,  
 nrow = 2,  
 labels = LETTERS[1:4]) #add letters A-D as figure labels

## Warning: Removed 2370 rows containing non-finite values (stat\_ydensity).

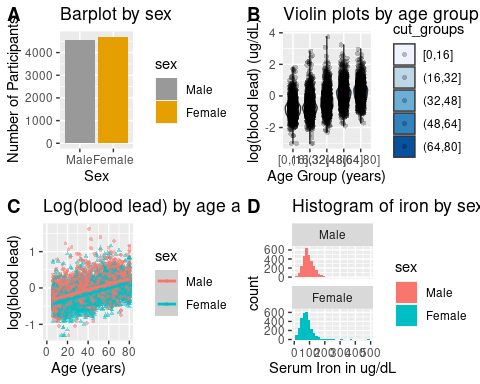
## Warning: Removed 2370 rows containing missing values (geom\_point).

## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 2370 rows containing non-finite values (stat\_smooth).  
  
## Warning: Removed 2370 rows containing missing values (geom\_point).

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 3332 rows containing non-finite values (stat\_bin).



# Assign compiled plot object  
compiled\_plots <- ggarrange(bar\_plot, viol\_plot, scatter\_plot, facet\_plot,  
 ncol = 2,  
 nrow = 2,  
 labels = LETTERS[1:4])

## Warning: Removed 2370 rows containing non-finite values (stat\_ydensity).  
  
## Warning: Removed 2370 rows containing missing values (geom\_point).

## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 2370 rows containing non-finite values (stat\_smooth).  
  
## Warning: Removed 2370 rows containing missing values (geom\_point).

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 3332 rows containing non-finite values (stat\_bin).

# Save and export plots

# Recommended to save figures as both pdf for high-quality viewing and png for presentations  
  
# Save the plot as a pdf for viewing at a high resolution  
ggsave(filename = here("compiled\_nhanes\_plots.pdf"),  
 plot = compiled\_plots,  
 width = 14,  
 height = 9)  
  
# Save the plot as a png for presentation with lower resolution  
ggsave(filename = here("compiled\_nhanes\_plots.png"),  
 plot = compiled\_plots,  
 units = "in",  
 width = 14,  
 height = 9,  
 dpi = 300)