Exploratory Data Analysis with R

Interactive Data Visualization

Xuemao Zhang East Stroudsburg University

November 4, 2022

Outline

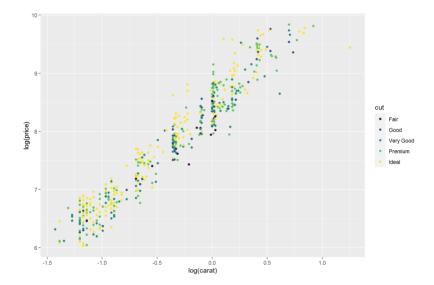
- Overview
- ggplot2 and plotly
- Data visualization with plotly
 - Pie/donut charts
 - Bars and histograms
 - Density and histogram overlay
 - Boxplots
 - Violin plots
 - Error bars
 - Scatter plot
 - 3d scatter plots
- Buttons
- You can disable/enable mouse click advance by pressing key 'k', when viewing the presentation.

Overview

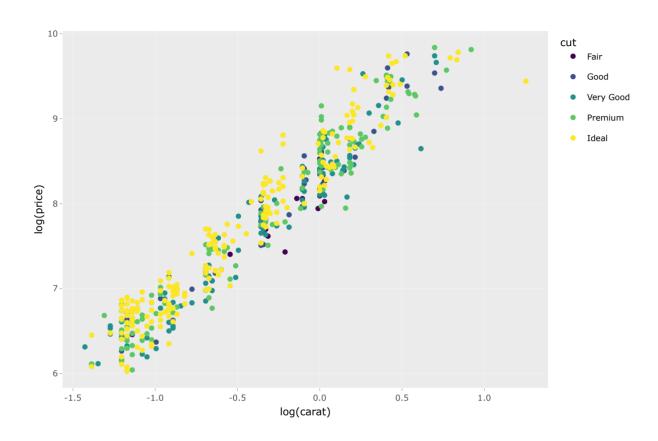
- Interactive graphics are different from animations in that the viewer has control.
- plotly: has come a long way in the last years, started as part of Carson Sievert's PhD thesis research. The beauty is that it builds directly onto ggplot2
- There are two main ways to creating a plotly object:
 - Transforming a ggplot2 object (via ggplotly()) into a plotly object
 - Directly initializing a plotly object with plot_ly()/plot_geo()/plot_mapbox(). They do not follow the grammar of graphics.

■ Calling plotly::ggplotly() after bringing a ggplot up will coerce it into an interactive plot. It may fail sometimes.

```
library(plotly);
library(dplyr);
library(ggplot2);
data(diamonds, package = "ggplot2");
set.seed(37);
diamonds1 = sample_n(diamonds, size=500);
p <- ggplot(diamonds1, aes(x = log(carat), y = log(price), color=cut))+
    geom_point();
p</pre>
```

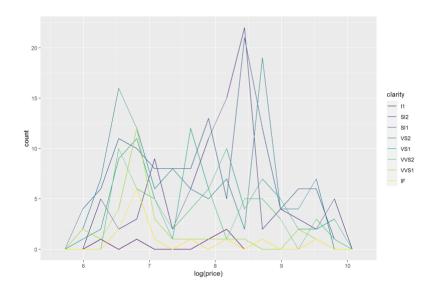


ggplotly(p);

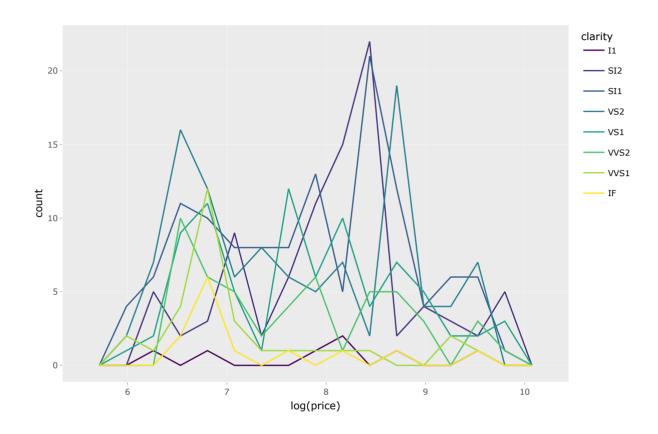


- ggplot2::geom_freqpoly() is used to visualise the distribution of a single continuous variable by dividing the x axis into bins and counting the number of observations in each bin
- geom_freqpoly() produces a frequency polygon for each level of that variable

```
p <- ggplot(diamonds1, aes(x = log(price), color = clarity)) +
    geom_freqpoly(bins=15);
p</pre>
```



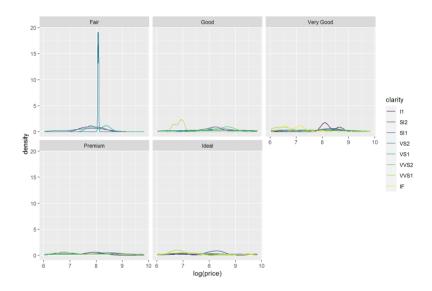
ggplotly(p);



■ stat = "density" is used to display relative rather than absolute frequencies.

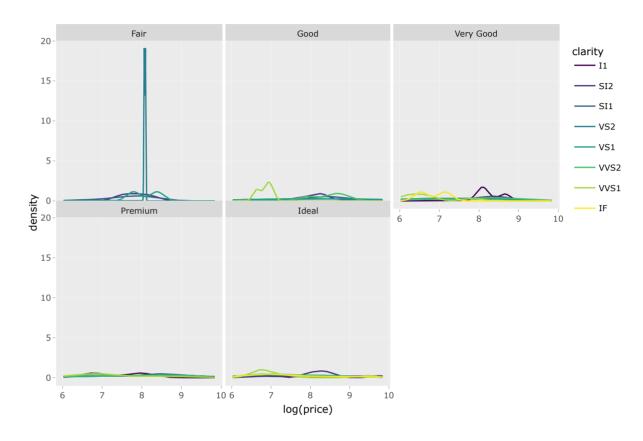
```
p <- ggplot(diamonds1, aes(x = log(price), color = clarity)) +
    geom_freqpoly(stat = "density") +
    facet_wrap(~cut);
p</pre>
```

Warning: Groups with fewer than two data points have been dropped.



ggplotly(p);

Warning: Groups with fewer than two data points have been dropped.



• Scatterplot(ly) matrix

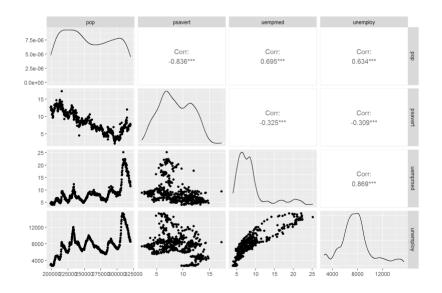
```
library(GGally);
data(economics, package = "ggplot2");
dim(economics);
```

```
## [1] 574 6
```

```
str(economics);
```

```
## spec_tbl_df [574 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ date : Date[1:574], format: "1967-07-01" "1967-08-01" ...
## $ pce : num [1:574] 507 510 516 512 517 ...
## $ pop : num [1:574] 198712 198911 199113 199311 199498 ...
## $ psavert : num [1:574] 12.6 12.6 11.9 12.9 12.8 11.8 11.7 12.3 11.7 12.3 ...
## $ uempmed : num [1:574] 4.5 4.7 4.6 4.9 4.7 4.8 5.1 4.5 4.1 4.6 ...
## $ unemploy: num [1:574] 2944 2945 2958 3143 3066 ...
```

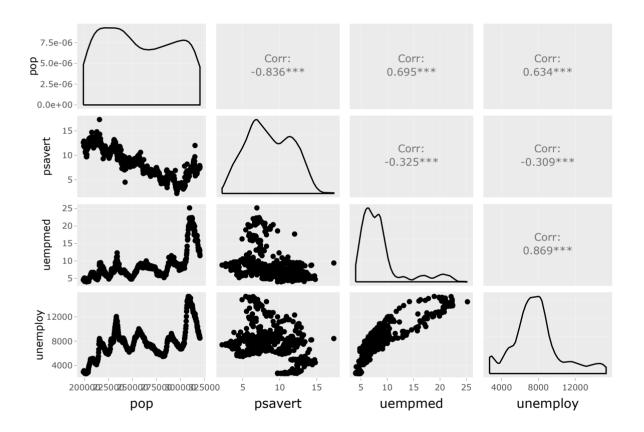
```
p <- ggpairs(economics[,3:6]);
p</pre>
```



• Scatterplot(ly) matrix

ggplotly(p)

```
## Warning: Can only have one: highlight
## Warning: Can only have one: highlight
## Warning: Can only have one: highlight
```



- Plotly is both a commercial service and open source product for creating high end interactive visualizations. The plotly package allows you to create plotly interactive graphs from within R.
- Any graph made with the plotly R package is powered by the JavaScript library plotly.js((MIT licensed)). The plot_ly() function provides a 'direct' interface to plotly.js with some additional abstractions to help reduce typing.
- The plot_ly() function has numerous arguments that are unique to the R package (e.g., color, stroke, span, symbol, linetype, etc) and make it easier to encode data variables.
 - This function maps R objects to plotly.js. It provides abstractions for doing common things (e.g. mapping data values to fill colors (via color) or creating animations (via frame)) and sets some different defaults to make the interface feel more 'R-like' (i.e., closer to plot() and ggplot2::gplot()).

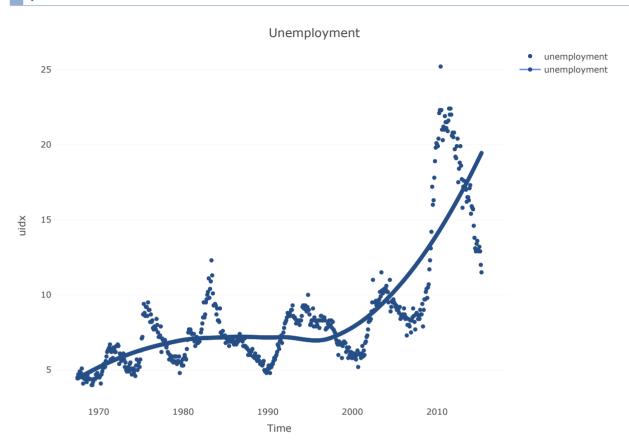
- A plotly.js figure contains one (or more) trace(s), and every trace has a type. The trace type scatter is great for drawing low-level geometries (e.g., points, lines, text, and polygons) and provides the foundation for many add_*() functions (e.g., add_markers(), add_lines(), add_paths(), add_segments(), add_ribbons(), add_area(), and add_polygons()) as well as many ggplotly() charts.
- The pipe operator %>% is used in the package.
- The types of traces are similar as geom_* in ggplot2. See R Figure Reference: image Traces.
- We follow the book Interactive web-based data visualization with R, plotly, and shiny to show some interactive data visualizations.

■ Plotly's graph description places attributes into two categories: traces (which describe a single series of data in a graph) and layout attributes that apply to the rest of the chart, like the title, xaxis, or annotations).

```
library(plotly)
p <- plot_ly(economics,</pre>
                                # all "scatter" attributes:
         type = "scatter",
         https://plotly.com/r/reference/#scatter
                             # the drawing mode for this scatter
         mode = "markers",
         trace
                                 # more about scatter's "x":
         x = \sim date
         /r/reference/#scatter-x
                                 # more about scatter's "y":
         y = \sim uempmed
         /r/reference/#scatter-y
         name = "unemployment", # more about scatter's "name":
         /r/reference/#scatter-name
         marker = list(
                                 # marker is a named list, valid keys:
         /r/reference/#scatter-marker
         color="#264E86"
                                # more about marker's "color" attribute:
         /r/reference/#scatter-marker-color
                                 # see https://htmlcolorcodes.com/ for
         colors
               )) %>%
  add_trace(x = ~date,
                                                                 #
         scatter's "x": /r/reference/#scatter-x
            y = ~fitted((loess(uempmed ~ as.numeric(date)))),
         scatter's "y": /r/reference/#scatter-y
            mode = 'lines+markers',
                                                                 #
         scatter's "y": /r/reference/#scatter-mode
            line = list(
                                                                # line is
         a named list, valid keys: /r/reference/#scatter-line
            color = "#5E88FC",
                                                                # line's
         "color": /r/reference/#scatter-line-color
            dash = "dashed"
                                                               # line's
         "dash" property: /r/reference/#scatter-line-dash
      ) %>%
```

```
layout(
                              # all of layout's properties:
      /r/reference/#layout
      title = "Unemployment", # layout's title: /r/reference/#layout-
      title
      xaxis = list(
                              # layout's xaxis is a named list. List
      of valid keys: /r/reference/#layout-xaxis
         title = "Time", # xaxis's title: /r/reference/#layout-
      xaxis-title
         showgrid = F),  # xaxis's showgrid:
      /r/reference/#layout-xaxis-showgrid
      yaxis = list(
                              # layout's yaxis is a named list. List
      of valid keys: /r/reference/#layout-yaxis
          title = "uidx") # yaxis's title: /r/reference/#layout-
      yaxis-title
    )
```

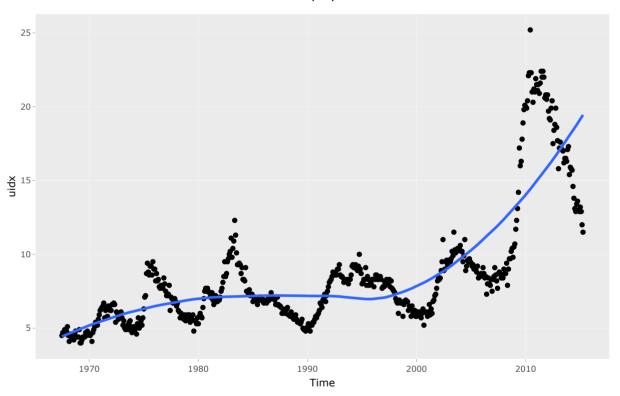




using ggplot2

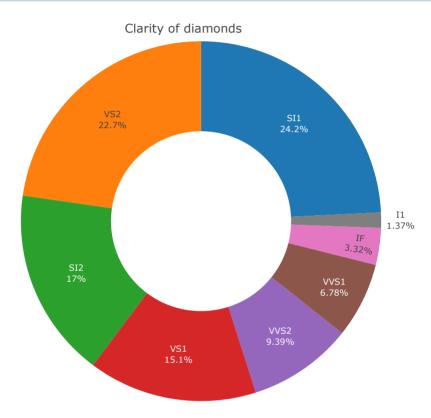
```
p=ggplot(data=economics, aes(x=date, y=uempmed) )+
    geom_point()+
    geom_smooth(formula=y~x, method="loess", se=F)+
    ggtitle("Unemployment")+
    theme(plot.title=element_text(hjust = 0.5))+#center the title
    labs(x="Time", y="uidx");
ggplotly(p)
```

Unemployment



Pie/donut charts

```
diamonds%>%group_by(clarity)%>%summarize(freq = n())%>% #
  plot_ly(labels = ~clarity, values = ~freq)%>%
  add_pie(hole = 0.5, text=~clarity)%>%
  layout(title = "Clarity of diamonds", showlegend =F);
```

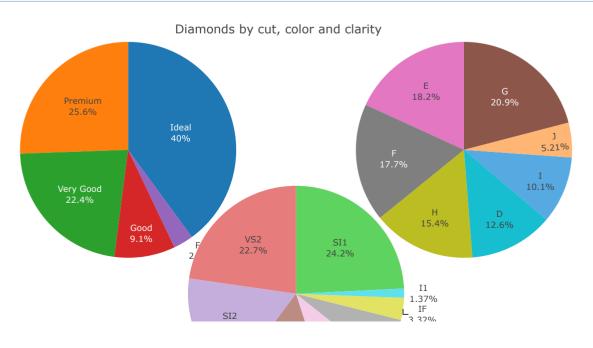


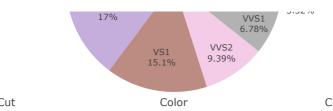
■ Polar coordinates are not yet supported by ggplotly(). So we cannot make a ggplot2 pie chart interactive.

Pie/donut charts

Pie charts with subplots

```
diamonds%>%group by(cut)%>%summarize(freq = n())->data1;
diamonds%>%group by(color)%>%summarize(freq = n())->data2;
diamonds%>%group_by(clarity)%>%summarize(freq = n())->data3;
fig <- plot_ly();</pre>
fig<-fig %>%add pie(data=data1, labels = ~cut, values = ~freq,
         text=~cut, domain = list(x = c(0, 0.4), y = c(0.4, 1));
fig<-fig %>%add pie(data=data2, labels = ~color, values =
         \simfreq,text=\simcolor, domain = list(x = c(0.6, 1), y = c(0.4,
         1)));
fig<-fig %>%add pie(data=data3, labels = ~clarity, values = ~freq,
         text=\simclarity, domain = list(x = c(0.25, 0.75), y = c(0,
         0.6)));
fig %>% layout(title = "Diamonds by cut, color and clarity", showlegend
  add_annotations( x=c(0.2, 0.5, 0.8), y=-0.05, text = c("Cut", 0.5, 0.8)
         "Color", "Clarity"),
    font = list(size = 15),
    xref = "paper", yref = "paper", xanchor = "center", showarrow =
         FALSE); # https://plotly.com/r/reference/layout/annotations/
```



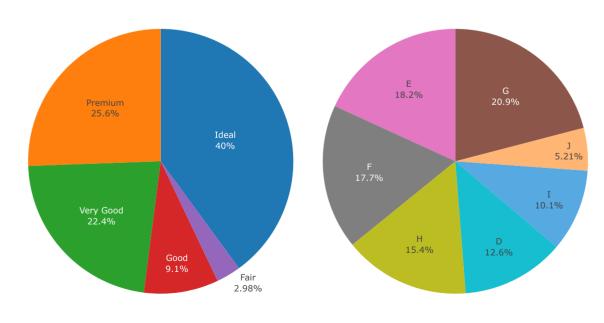


Clarity Cut

Pie/donut charts

Subplots Using Grid

Diamonds by cut and color



Cut Color

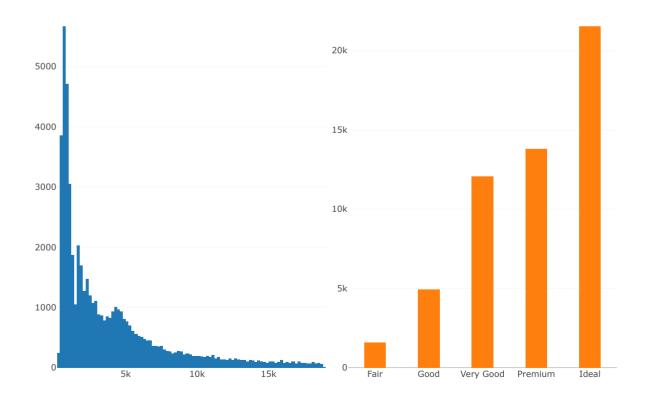
Bars and histograms

- The add_bars() and add_histogram() functions wrap the bar and histogram plotly.js trace types.
- The subplot() function provides a flexible interface for merging plotly objects into a single object (i.e., view).

```
library(dplyr)
p1 <- plot_ly(diamonds, x = ~price) %>% add_histogram();

p2 <- diamonds %>% group_by(cut)%>%summarise(freq=n())%>%
    plot_ly(x = ~cut, y = ~freq)%>%add_bars(width = 0.4);

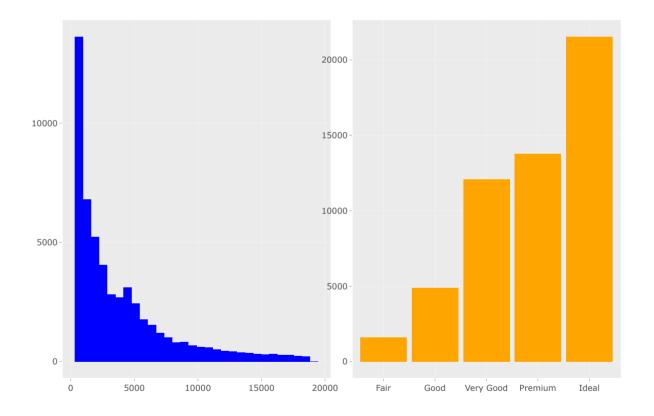
fig <- subplot(p1, p2);
fig%>%hide_legend(); #removing the Legend due to the merging
```



Bars and histograms

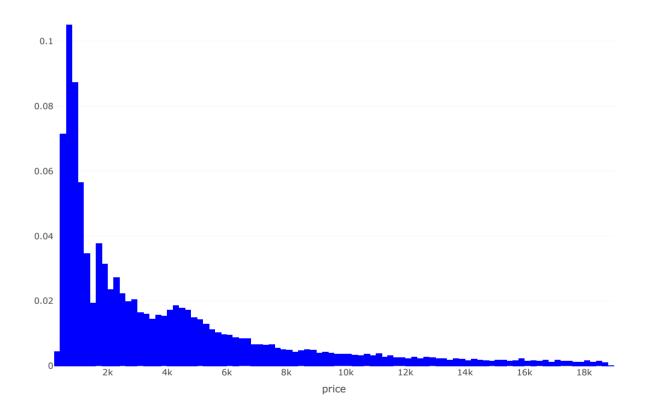
using ggplot2

```
p1 <- ggplot(diamonds, aes(x = price)) +
    geom_histogram(fill='blue');
p2 <- ggplot(diamonds, aes(x = cut))+
    geom_bar(fill="orange");
ply1=ggplotly(p1)
ply2=ggplotly(p2)
subplot(ply1, ply2)</pre>
```



Bars and histograms

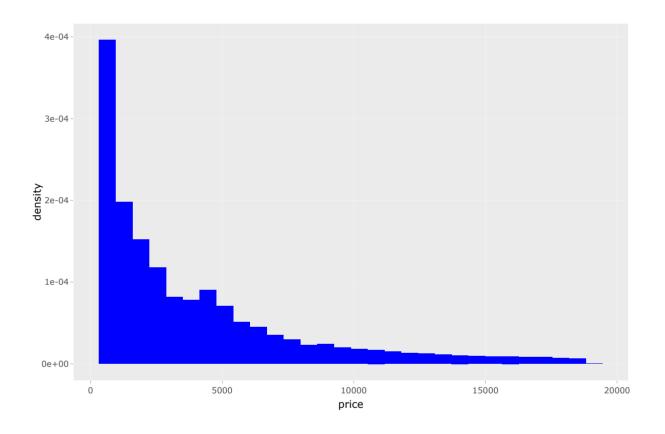
Probability (relative frequency) histogram



Bars and histograms

using ggplot2

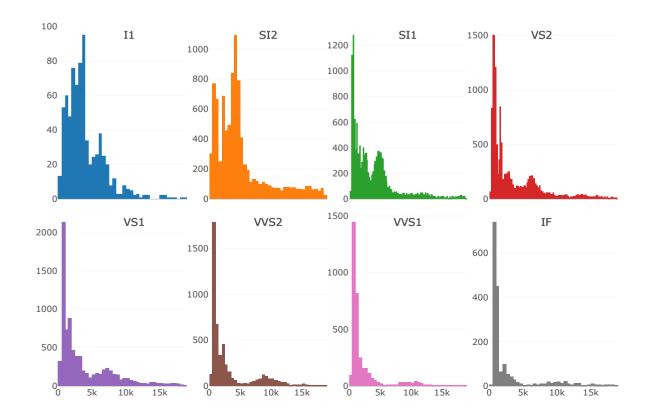
```
p=ggplot(diamonds, aes(x = price)) +
  geom_histogram(aes(y= ..density..), fill='blue');
ggplotly(p)
```



Bars and histograms

 display of diamond price by diamond clarity (like faceting method in ggplot2)

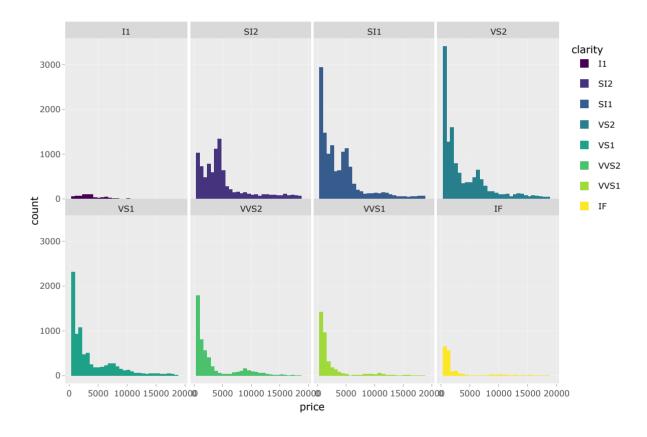
```
one_plot <- function(d){
  plot_ly(d, x = ~price) %>%
    add_histogram()%>%
    add_annotations(text=~unique(clarity), x = 0.5, y = 1,
        xref = "paper", yref = "paper", xanchor = "middle",
    yanchor = "top",showarrow = FALSE,
    font = list(size = 15, face="bold") )
# https://plotly.com/r/reference/layout/annotations/
  }
```



Bars and histograms

display of diamond price by diamond clarity (using ggplot2)

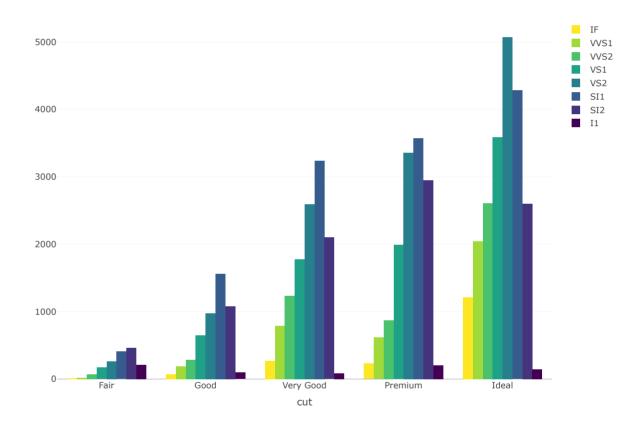
```
p=ggplot(diamonds, aes(x = price, fill=clarity)) +
   geom_histogram()+
   facet_wrap(~clarity, nrow=2, ncol=4);
ggplotly(p);
```



Bars and histograms

Grouped Bar Charts

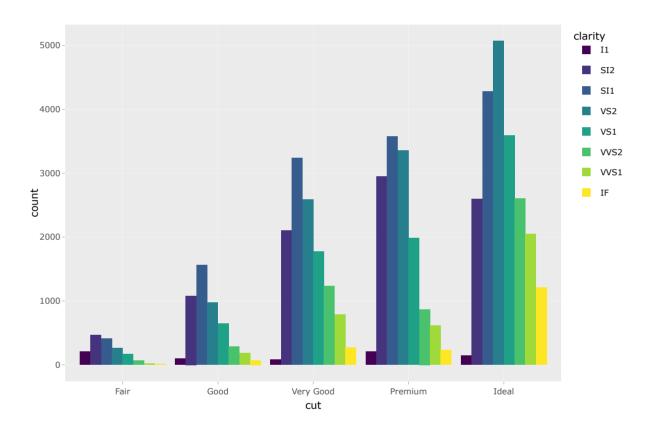
```
plot_ly(diamonds, x = ~cut, color = ~clarity) %>%
  add_histogram(); #both variables are categorical
```



Bars and histograms

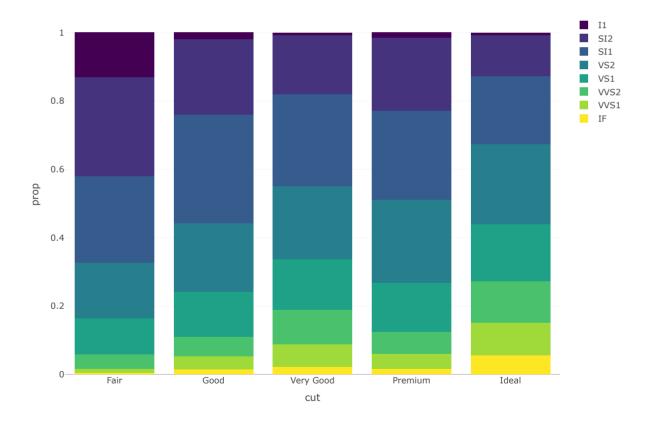
Grouped Bar Charts using ggplot2

```
p=ggplot(diamonds, aes(x = cut, fill = clarity)) +
  geom_bar(position="dodge", stat = "count"); #position="stack"
  ggplotly(p)
```



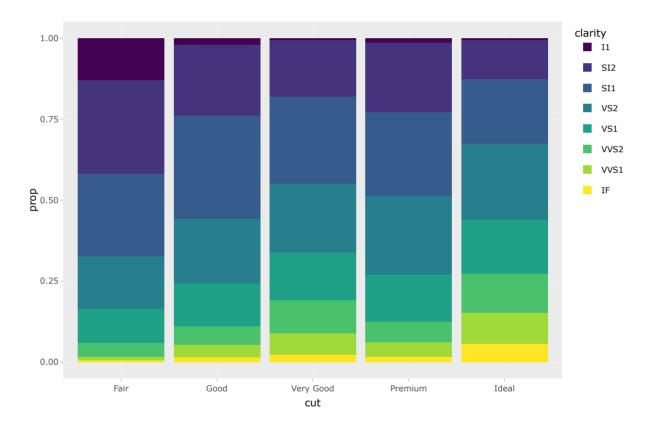
Bars and histograms

- Grouped Bar Charts
 - we show the relative frequency of diamonds by clarity, given a cut.

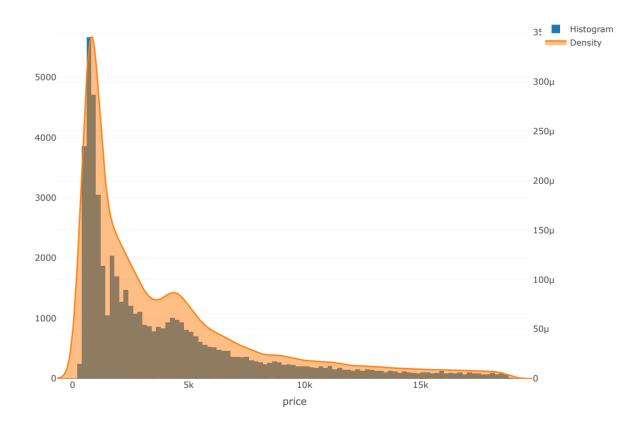


Bars and histograms

Grouped Bar Charts using ggplot2 (relative frequency chart)



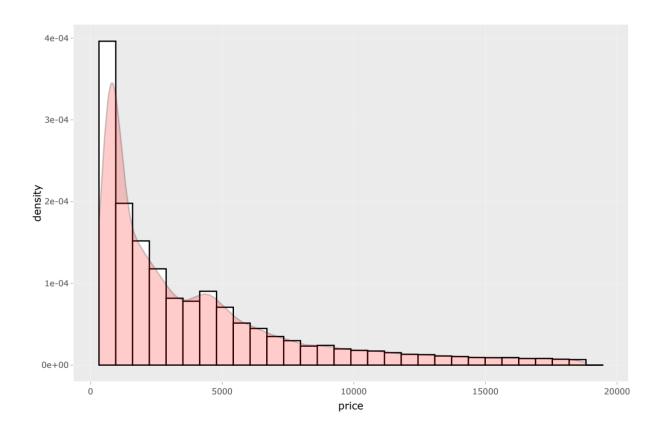
Density and histogram overlay



Density and histogram overlay

using ggplot2

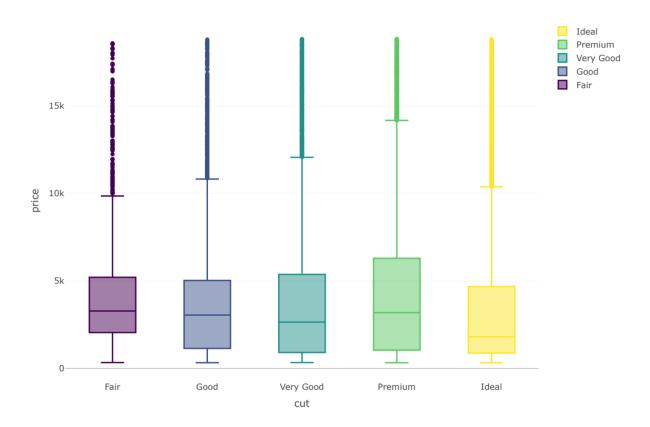
```
p=ggplot(data=diamonds, aes(x=price)) +
geom_histogram(aes(y= ..density..), color="black",fill="white")+
#Histogram with density instead of count on y-axis
geom_density(alpha=.2, fill="red");
ggplotly(p)
```



Boxplots

■ Boxplots of price by factor cut

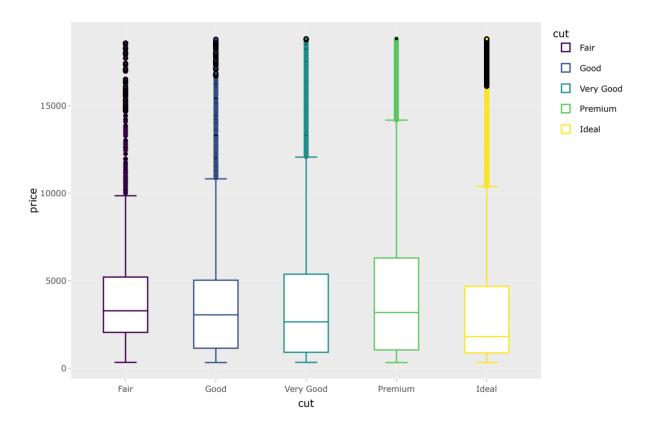
```
plot_ly(diamonds, x=~cut, y=~price, color=~cut)%>%
  add_boxplot()
```



Boxplots

■ Boxplots of price by factor cut using ggplot2

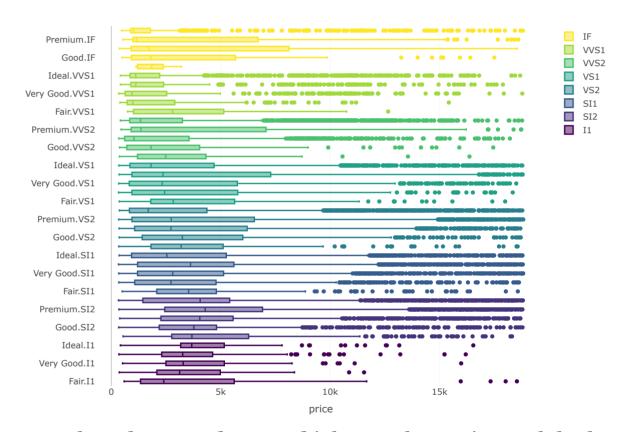
```
p=ggplot(diamonds, aes(x=cut, y=price, color=cut))+
  geom_boxplot();
ggplotly(p)
```



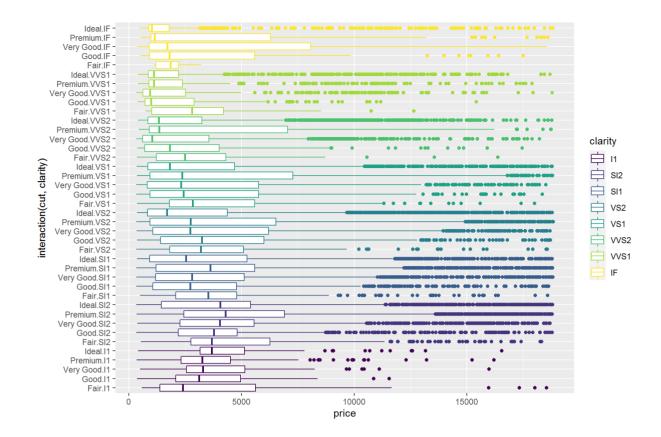
Boxplots

Boxplots of price by factor cut and clarity

```
plot_ly(diamonds, x = ~price, y = ~interaction(cut, clarity)) %>%
  add_boxplot(color = ~clarity)%>%
  layout(yaxis = list(title = "")) #remove the ylab
```



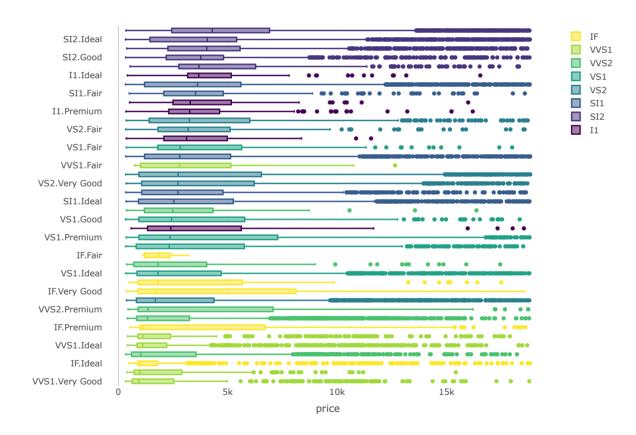
ggplot2 does not do a good job. But the static graph looks good



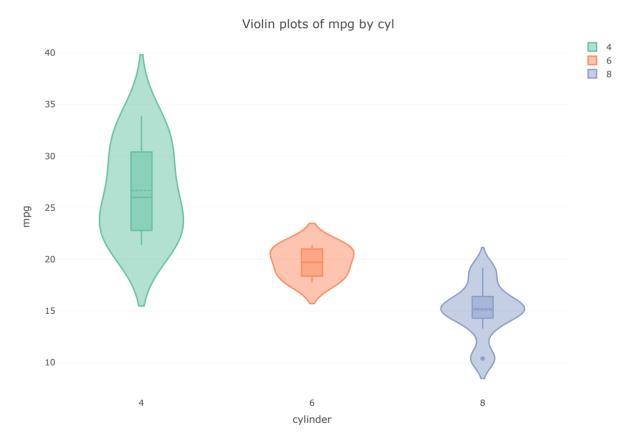
#ggplotly(p)

Boxplots

Sort the boxplots by medians



Violin plots

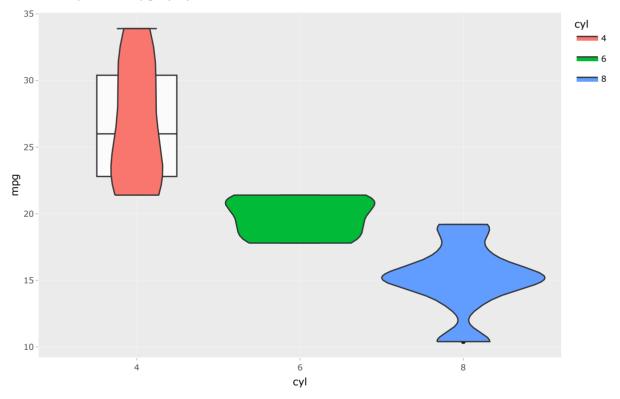


Violin plots

- using ggplot2
 - Not all boxes are shown 😥

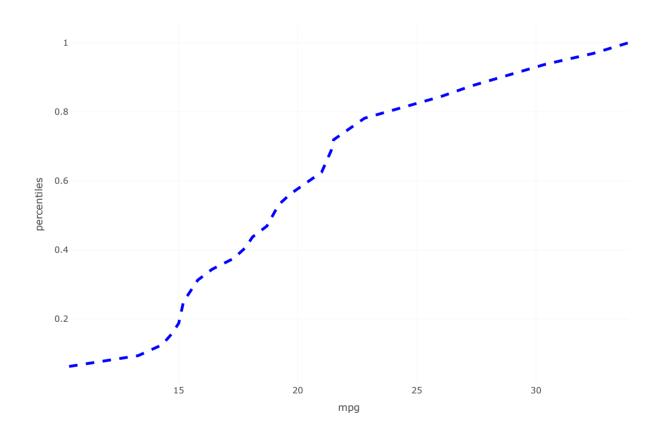
```
p=ggplot(data=mtcars, aes(x=cyl,y=mpg) ) +
   geom_violin(aes(fill=cyl))+
   geom_boxplot(width = 0.1, alpha=0.8)+
   labs(title = "Violin plots of mpg by cyl");
ggplotly(p);
```

Violin plots of mpg by cyl



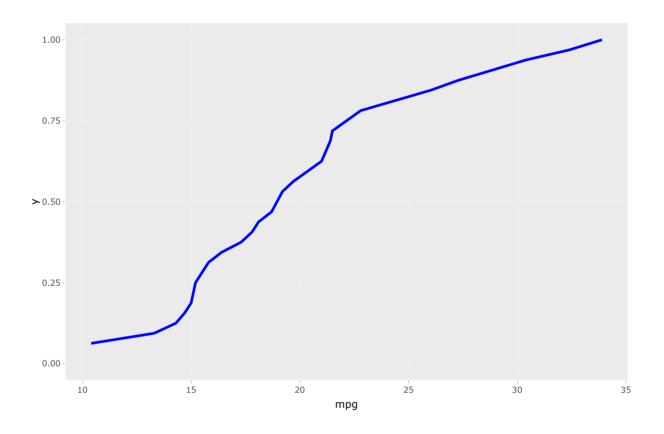
Cumulative Frequencies

```
mtcars %>%arrange(mpg)->mtcars1;
Fn=ecdf(mtcars1$mpg); # ecdf returns a *function*
mtcars1<-mtcars1%>%mutate(percentiles=Fn(mpg));
#Fn(mtcars1$mpg) #returns the percentiles
mtcars1%>%plot_ly(x = ~mpg, y = ~percentiles, type = 'scatter', mode = 'lines', name="cdf",
    line = list(width = 4, color="blue",dash = "dash"));
```

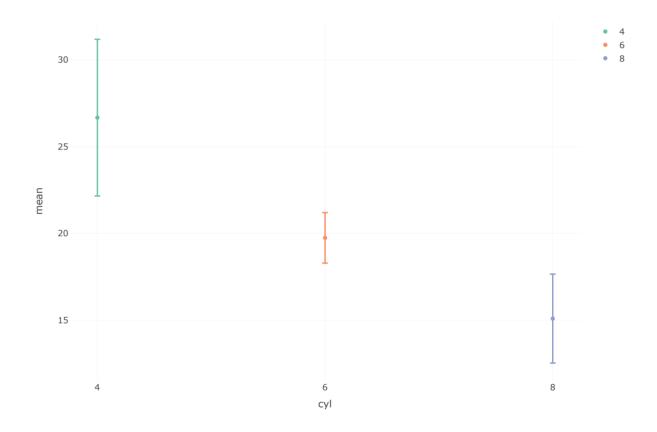


Cumulative Frequencies using ggplot2

```
p=ggplot(mtcars, aes(x=mpg)) +
stat_ecdf(geom = "line", color="blue", size=1);
ggplotly(p);
```



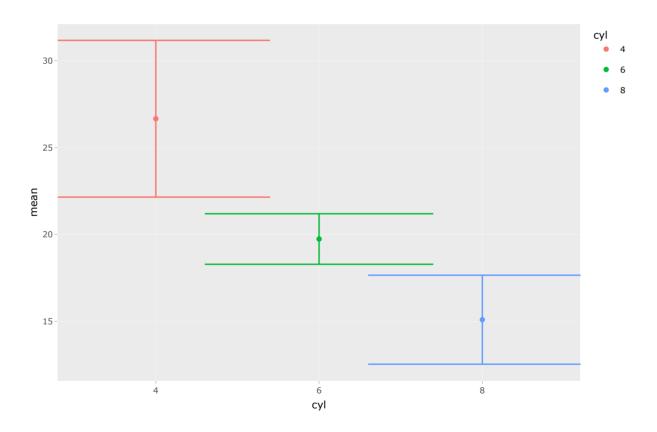
Error bars



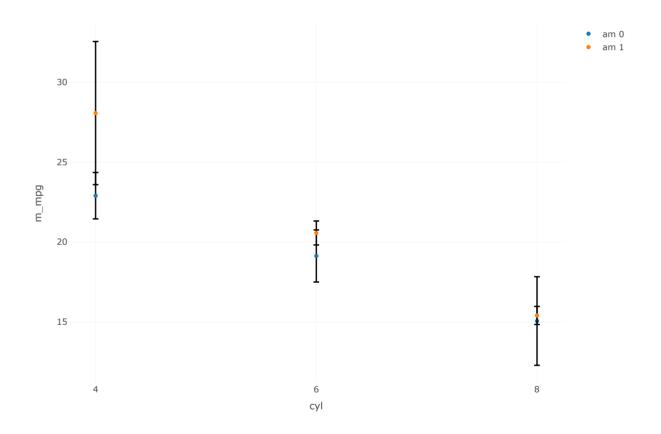
Error bars

using ggplot2

```
p<-mtcars%>%mutate(cyl=factor(cyl))%>%
  group_by(cyl)%>%summarise(mean=mean(mpg),sd=sd(mpg))%>%
  ggplot(aes(x=cyl,y=mean, color=cyl)) +
  geom_point()+
  geom_errorbar(aes(ymin=mean-sd, ymax=mean+sd));
ggplotly(p);
```



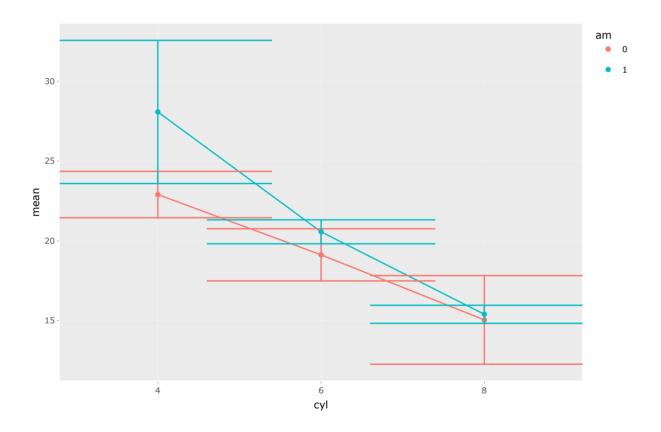
Error bars



Error bars

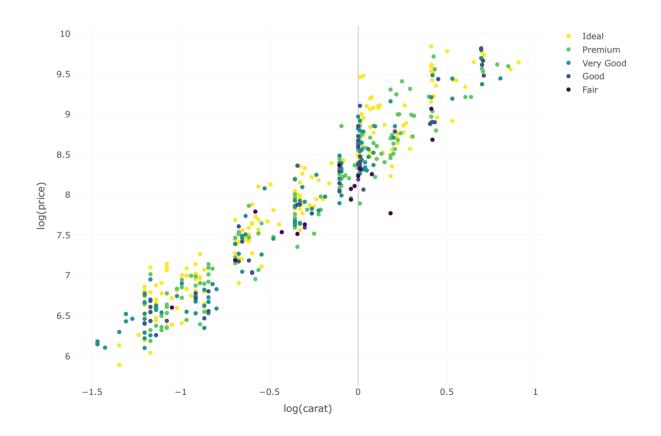
using ggplot2

```
p=mtcars%>%mutate(cyl=factor(cyl), am=factor(am))%>%
    group_by(cyl,am)%>%summarise(mean=mean(mpg),sd=sd(mpg))%>%
    ggplot(aes(x=cyl,y=mean, color=am)) + geom_point()+
geom_line(aes(group=am))+
geom_errorbar(aes(ymin=mean-sd, ymax=mean+sd));
ggplotly(p);
```



Scatter plot

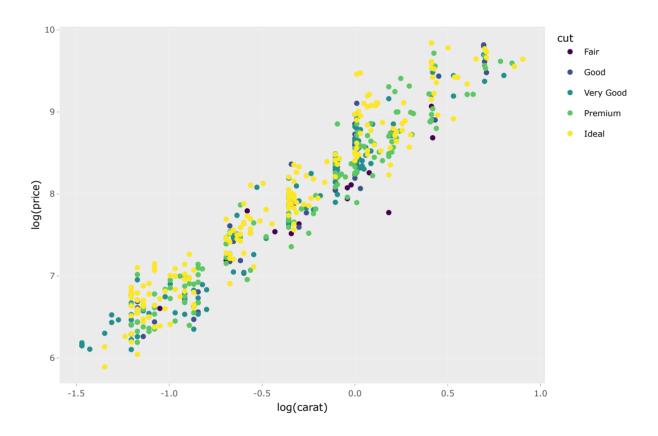
- For scatter plot matrix, please see https://plotly.com/r/splom/
- Please see our first plotly example



Scatter plot using ggplot2

We have seen this

```
p <- ggplot(diamonds2, aes(x = log(carat), y = log(price), color=cut))+
  geom_point();
ggplotly(p);</pre>
```



Scatter plot

• 3d scatter plots

```
mtcars%>%mutate(cyl=factor(cyl))%>%
  plot_ly(x = ~mpg, y = ~disp, z = ~cyl) %>%
  add_markers(color = ~cyl)
```

• 4

3d scatter plots

```
mtcars%>%mutate(cyl=factor(cyl))%>%
  plot_ly(x = ~mpg, y = ~disp, z = ~wt,color = ~cyl) %>%
  add_markers();
```

468

3d scatter plots

```
mtcars%>%mutate(cyl=factor(cyl))%>%
  plot_ly(x = ~mpg, y = ~disp, z = ~wt, color = ~cyl) %>%
  add_markers()%>%add_lines();
```

```
4
6
8
4
6
8
```

Buttons

Dropdown events

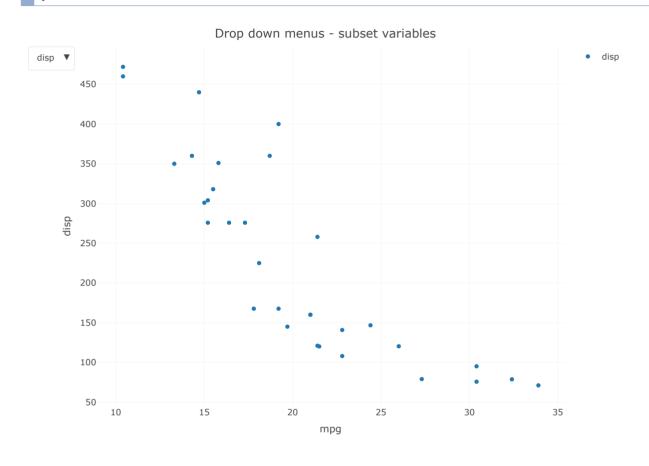
- The updatemenu method determines which plotly.js function will be used to modify the chart. There are 4 possible methods:
 - restyle: modify data or data attributes
 - update: modify data and layout attributes
 - relayout: modify layout attributes
 - animate: start or pause an animation (only available offline)

```
p=mtcars%>%plot ly(x = ~mpg, y = ~disp,
            name='disp', type='scatter', mode='markers') %>%
 add trace(y = ~hp, name = 'hp', type='scatter', mode='markers',
         visible=FALSE) %>%
 add trace(y = ~wt, name = 'wt', type='scatter', mode='markers',
         visible=FALSE) %>%
layout(
  title = "Drop down menus - subset variables",
  yaxis = list(title = "disp"),
            updatemenus = list(
            list(
              type= 'dropdownlist',
          buttons = list(
          list(method = "update",
               args = list(list(visible = list(TRUE, FALSE, FALSE)),
                           list(yaxis = list(title = "disp"))),
               label = "disp"),
          list(method = "update",
               args = list(list(visible = list(FALSE, TRUE, FALSE)),
                           list(yaxis = list(title = "hp"))),
               label = "hp"),
          list(method = "update",
               args = list(list(visible = list(FALSE, FALSE, TRUE)),
                           list(yaxis = list(title = "wt"))),
               label = "wt")
          )
       )
      )
    )
```

Buttons

Dropdown events

p



Buttons

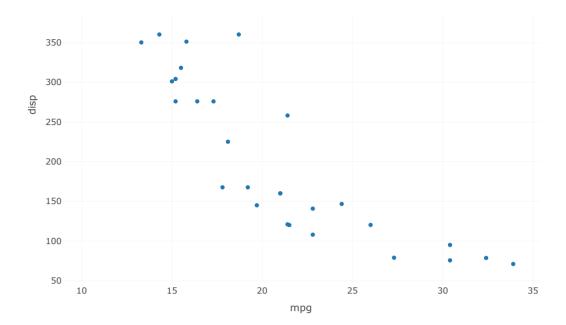
Dropdown events

- Try to change type type= 'buttons',
- Read https://plotly.com/r/custom-buttons/ for more.

```
mtcars%>%plot ly(x = ~mpg, y = ~disp,
            name='disp', type='scatter', mode='markers') %>%
 add_trace(y = ~hp, name = 'hp', type='scatter', mode='markers',
         visible=FALSE) %>%
  add trace(y = ~wt, name = 'wt', type='scatter', mode='markers',
         visible=FALSE) %>%
 layout(
   title = "subset variables",
  yaxis = list(title = "disp"),
            updatemenus = list(
            list(
              type= 'buttons',
          buttons = list(
          list(method = "update",
               args = list(list(visible = list(TRUE, FALSE, FALSE)),
                           list(vaxis = list(title = "disp"))),
               label = "disp"),
          list(method = "update",
               args = list(list(visible = list(FALSE, TRUE, FALSE)),
                           list(yaxis = list(title = "hp"))),
               label = "hp"),
          list(method = "update",
               args = list(list(visible = list(FALSE, FALSE, TRUE)),
                           list(vaxis = list(title = "wt"))),
               label = "wt")
          )
       )
      )
    )
```

disp hp 450 wt

subset variables



License



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.