CSM_2025 - Good, bad and ugly

Andrew Hooyman

2025-01-01

Data Visualization for communication

The purpose of a data visualization in a scientific article should be to convey 1. Measures of Central Tendency 2. Measures of Dispersion 3. Linear Trends

```
#install.packages(c("cowplot", "ggplot2", "ggpubr", "gghalves", "gtsummary", "curl"))


library(cowplot)
```

```
## Warning: package 'cowplot' was built under R version 4.0.4
```

```
library(ggplot2)
library(ggpubr)
```

```
##
## Attaching package: 'ggpubr'
```

```
## The following object is masked from 'package:cowplot':
##
##     get_legend
```

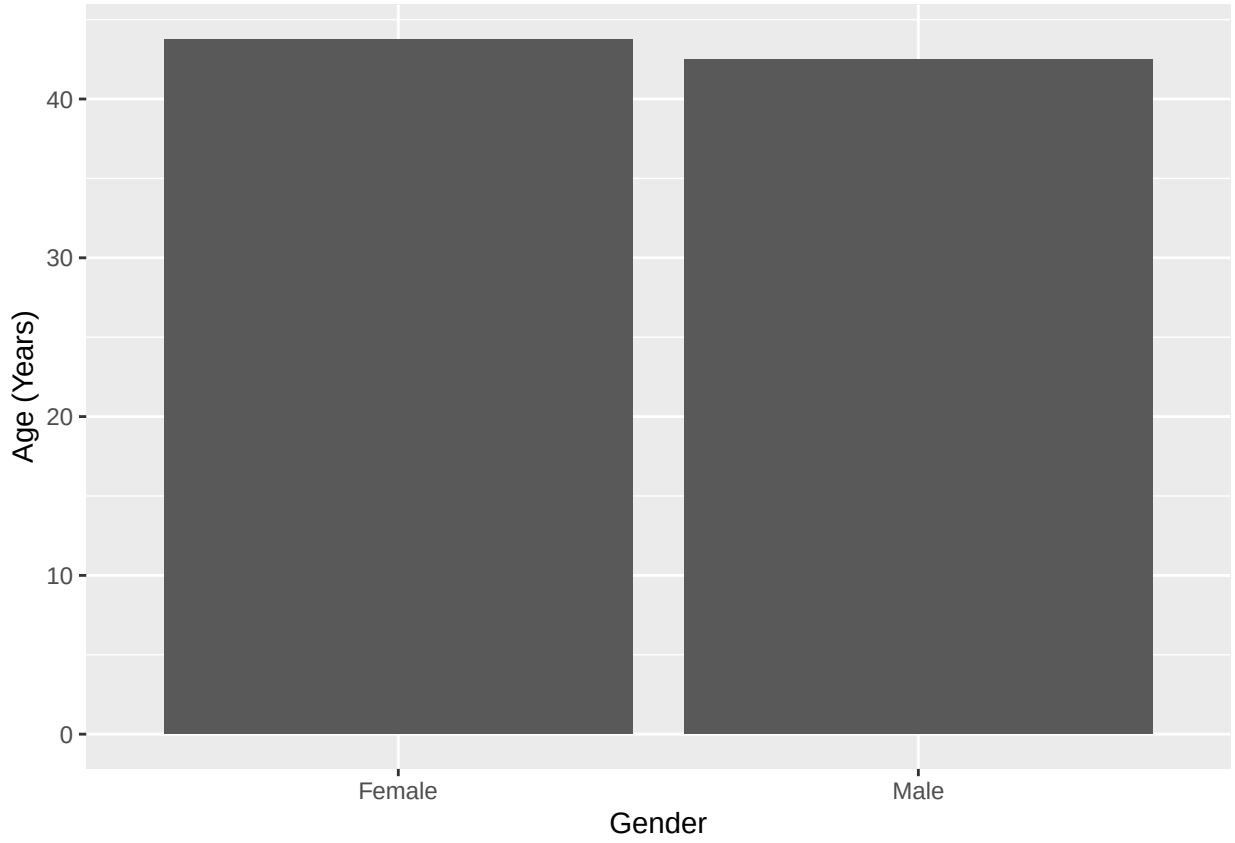
```
library(gghalves)
library(gtsummary)
library(curl)


url <- "https://raw.githubusercontent.com/hooymana/CSM_Data-Visualization_2025/main/healthcare-dataset-s
data <- read.csv(curl(url),na.strings = "N/A")

tbl_summary(data[,-1],
            statistic = list(
              all_continuous() ~ "{mean} ({sd})",
                 all_categorical() ~ "{n} ({p}%)"),
            digits = all_continuous() ~ 2,
               missing = "ifany"
            ) %>%
  bold_labels()
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**N = 5,109**
gender	
Female	2,994 (59%)
Male	2,115 (41%)
age	43.23 (22.61)
hypertension	498 (9.7%)
heart_disease	276 (5.4%)
ever_married	3,353 (66%)
work_type	
children	687 (13%)
Govt_job	657 (13%)
Never_worked	22 (0.4%)
Private	2,924 (57%)
Self-employed	819 (16%)
Residence_type	
Rural	2,513 (49%)
Urban	2,596 (51%)
avg_glucose_level	106.14 (45.29)
bmi	28.89 (7.85)
Unknown	201
smoking_status	
formerly smoked	884 (17%)
never smoked	1,892 (37%)
smokes	789 (15%)
Unknown	1,544 (30%)
stroke	249 (4.9%)

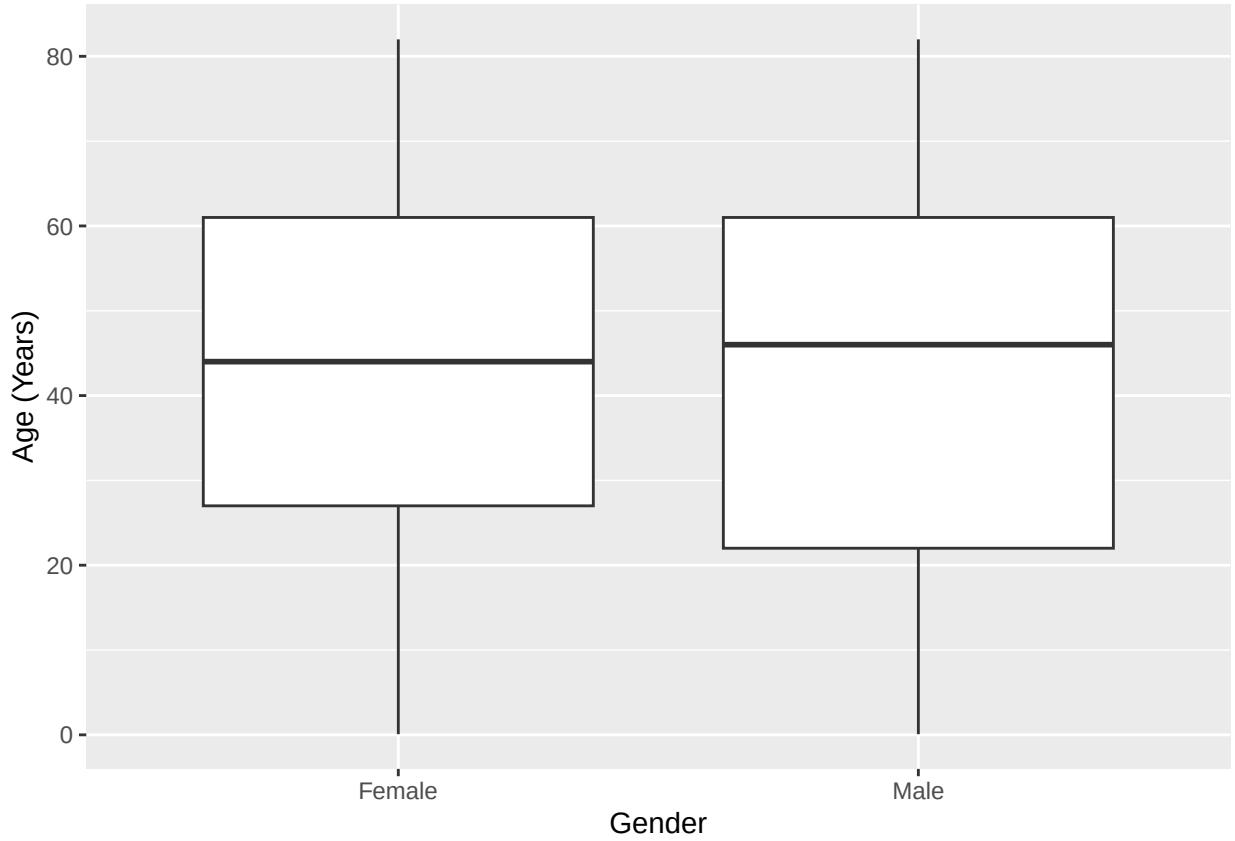
Graph from "A Simple Start" Slide

```
#Basic Bar Graph
ggplot(data,aes(x=gender,y=age))+
  stat_summary(fun="mean",geom="bar")+
  xlab("Gender")+
  ylab("Age (Years)")
```



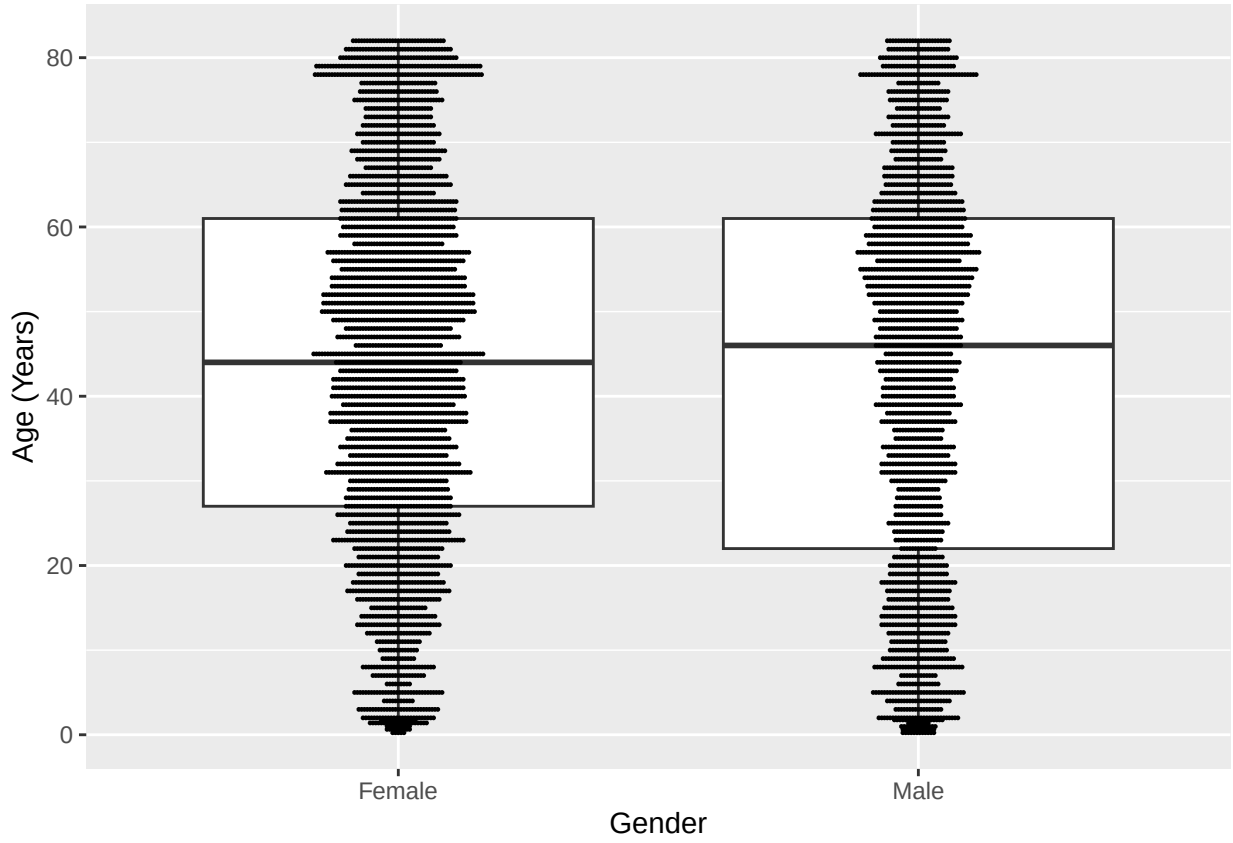
Graph from "Box it Out" Slide

```
#Boxplot
ggplot(data,aes(x=gender,y=age))+
  geom_boxplot()+
  xlab("Gender")+
  ylab("Age (Years)")
```



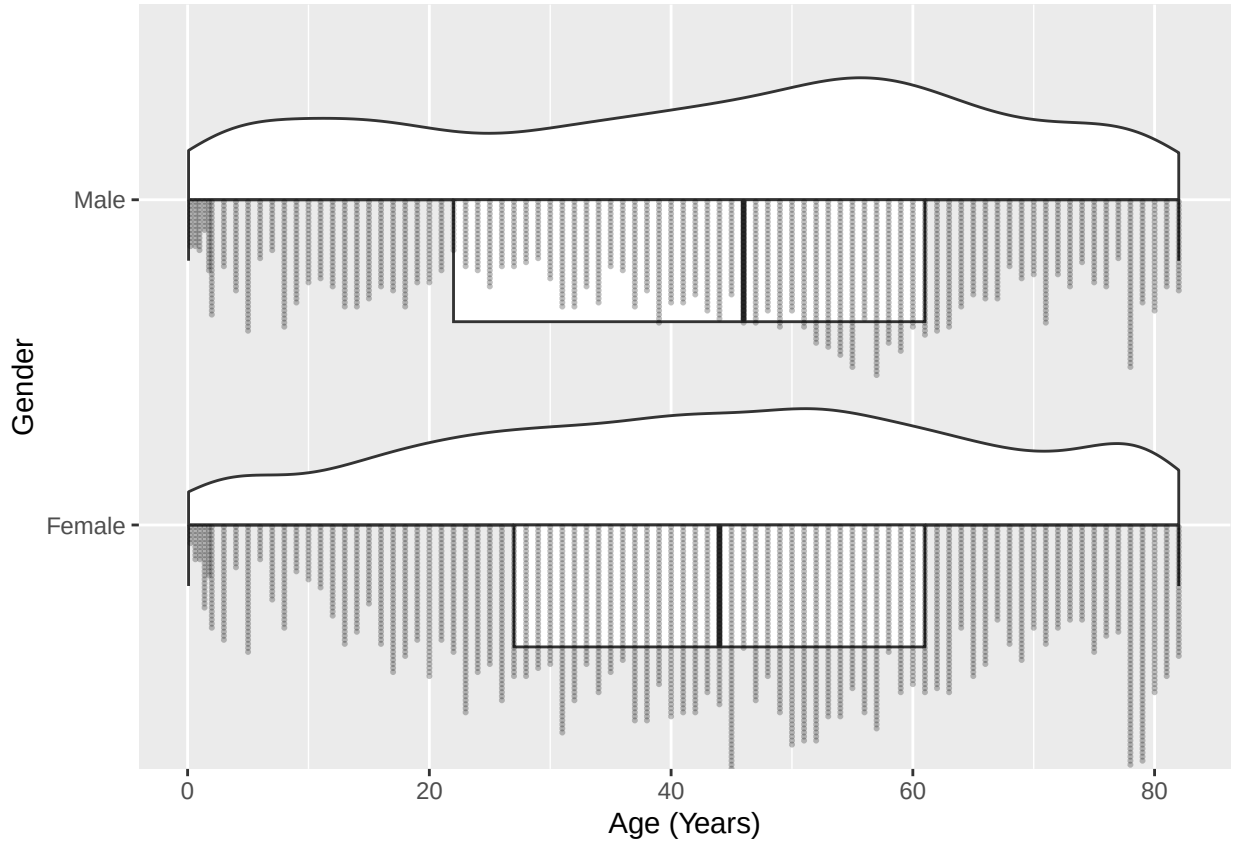
Graph from "Layer it up" Slide

```
#Boxplot + dotplot
ggplot(data,aes(x=gender,y=age))+
  geom_boxplot()+
  geom_dotplot(stackdir = "center",binaxis = "y",binwidth = 1/3)+
  xlab("Gender")+
  ylab("Age (Years)")
```



Graph from "Flip it out" Slide

```
#Rain cloud plot
ggplot(data,aes(x=gender,y=age))+
  #geom_half_boxplot(side="l",errorbar.length = 1)+
  geom_half_violin(side="r")+
  geom_half_boxplot(side="l")+
  geom_half_dotplot(stackdir = "down",binwidth=1/3,alpha=.25)+
  xlab("Gender")+
  ylab("Age (Years)")+
  coord_flip()
```

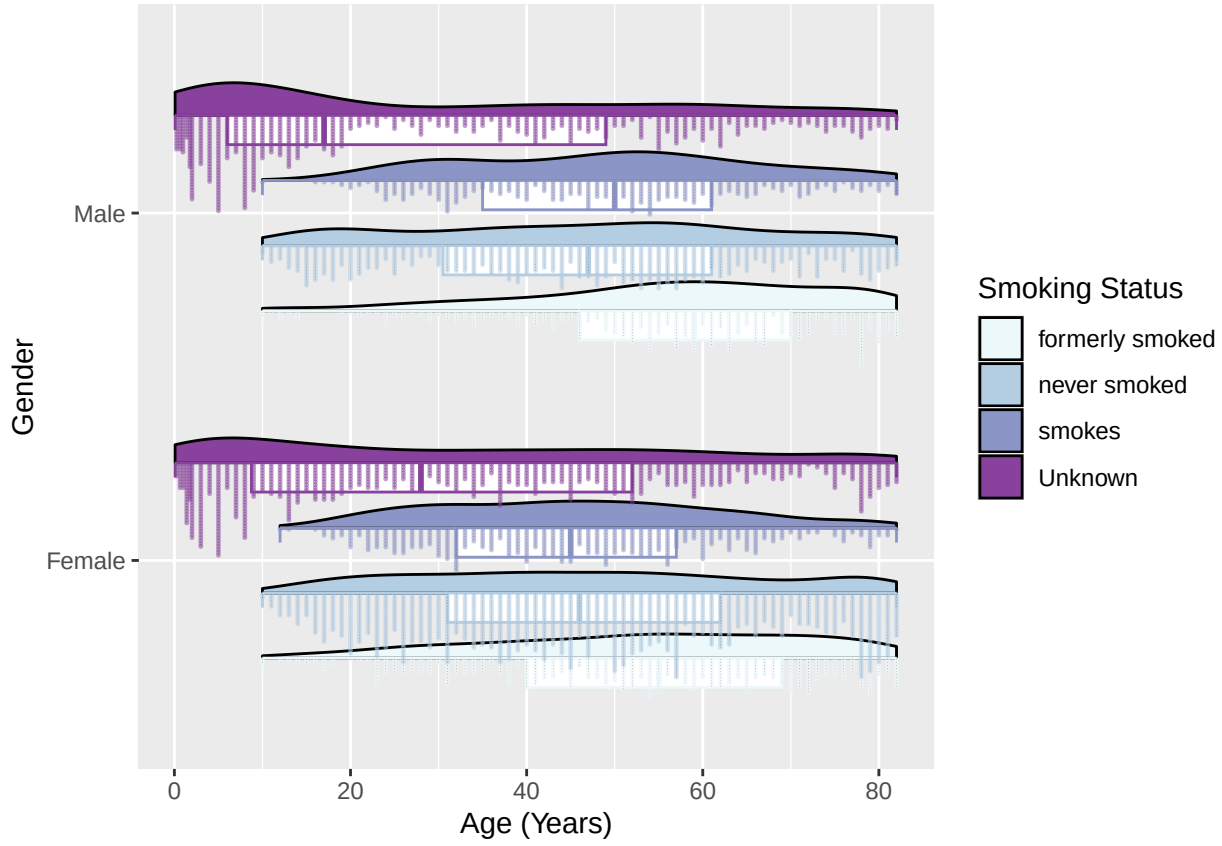


Graph from "Visual of Sub-groups" Slide

```
#Rain cloud plot, stratified by smoking status (color)
library(RColorBrewer) #customize your colors
```

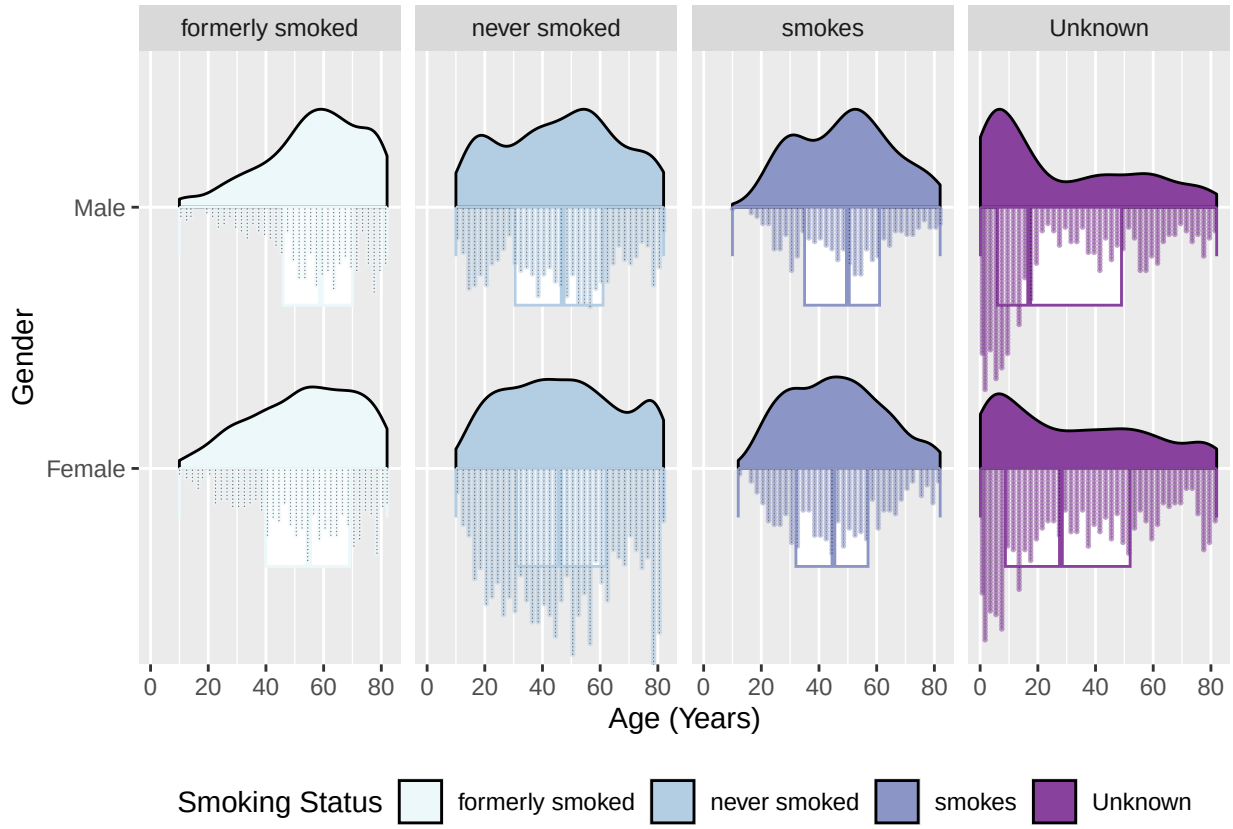
```
## Warning: package 'RColorBrewer' was built under R version 4.0.5
```

```
ggplot(data,aes(x=gender,y=age))+
  geom_half_violin(side="r",aes(fill=smoking_status),color="black")+
  geom_half_boxplot(side="l",show.legend = F,aes(color=smoking_status))+
  geom_half_dotplot(stackdir = "down",binwidth=1/3,alpha=.5,show.legend = F,aes(color=smoking_status))+
  scale_fill_manual(values = brewer.pal(4,"BuPu"))+
  scale_color_manual(values = brewer.pal(4,"BuPu"))+
  xlab("Gender")+
  ylab("Age (Years)")+
  coord_flip()+
  labs(fill="Smoking Status")
```



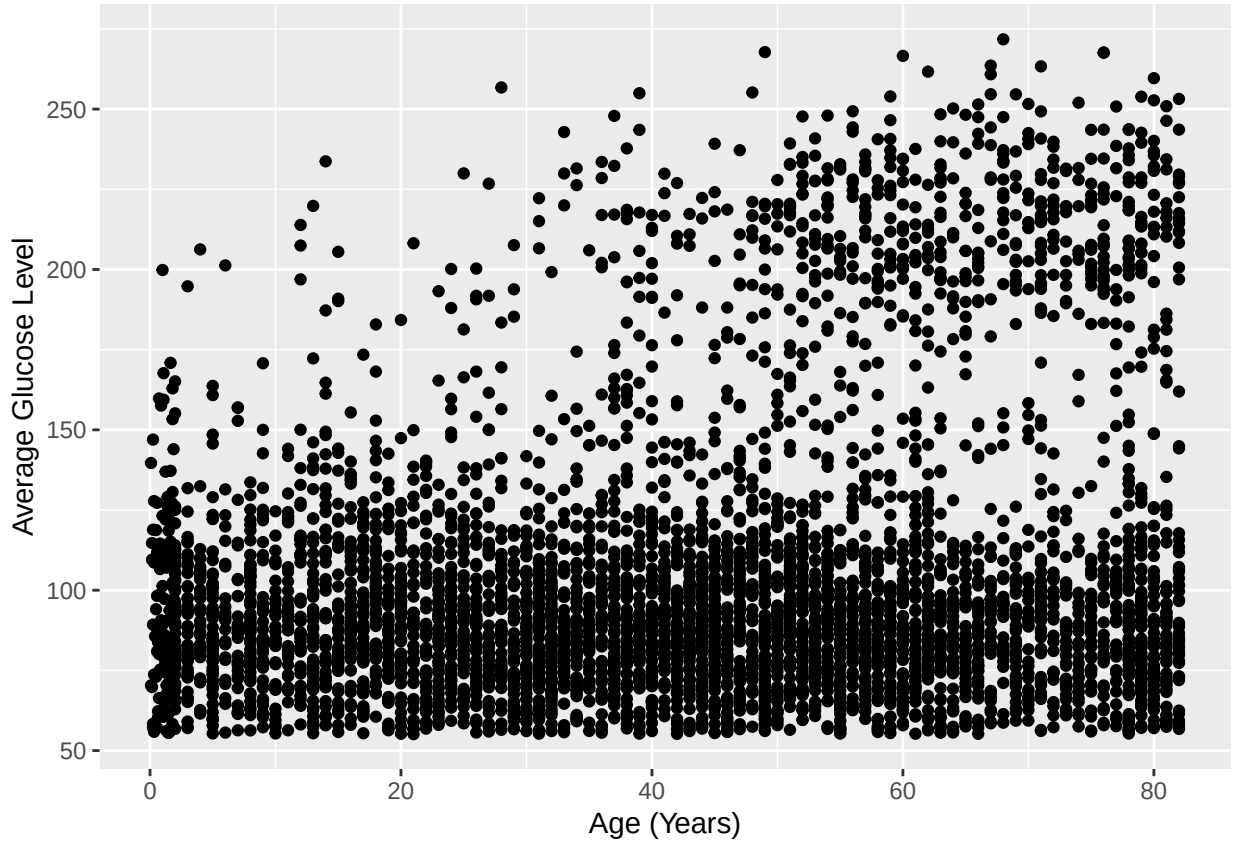
Graph from "Visual of Sub-groups" Slide 2

```
#Rain cloud plot, stratified by smoking status (color), coordinates flipped
ggplot(data,aes(x=gender,y=age,color=smoking_status))+
  geom_half_violin(side="r",aes(fill=smoking_status),color="black")+
  geom_half_boxplot(side="l",show.legend = F,aes(color=smoking_status))+
  geom_half_dotplot(stackdir = "down",binwidth=1.25,alpha=.5,show.legend = F,aes(color=smoking_status))+
  scale_fill_manual(values = brewer.pal(4,"BuPu"))+
  scale_color_manual(values = brewer.pal(4,"BuPu"))+
  xlab("Gender")+
  ylab("Age (Years)")+
  coord_flip()+
  facet_grid(~smoking_status)+
  theme(legend.position = "bottom")+
  labs(fill="Smoking Status")
```



Graph from "Switching Gears" Slide

```
#Basic Scatter plot
ggplot(data,aes(x=age,y=avg_glucose_level))+
  geom_point()+
  xlab("Age (Years)")+
  ylab("Average Glucose Level")
```

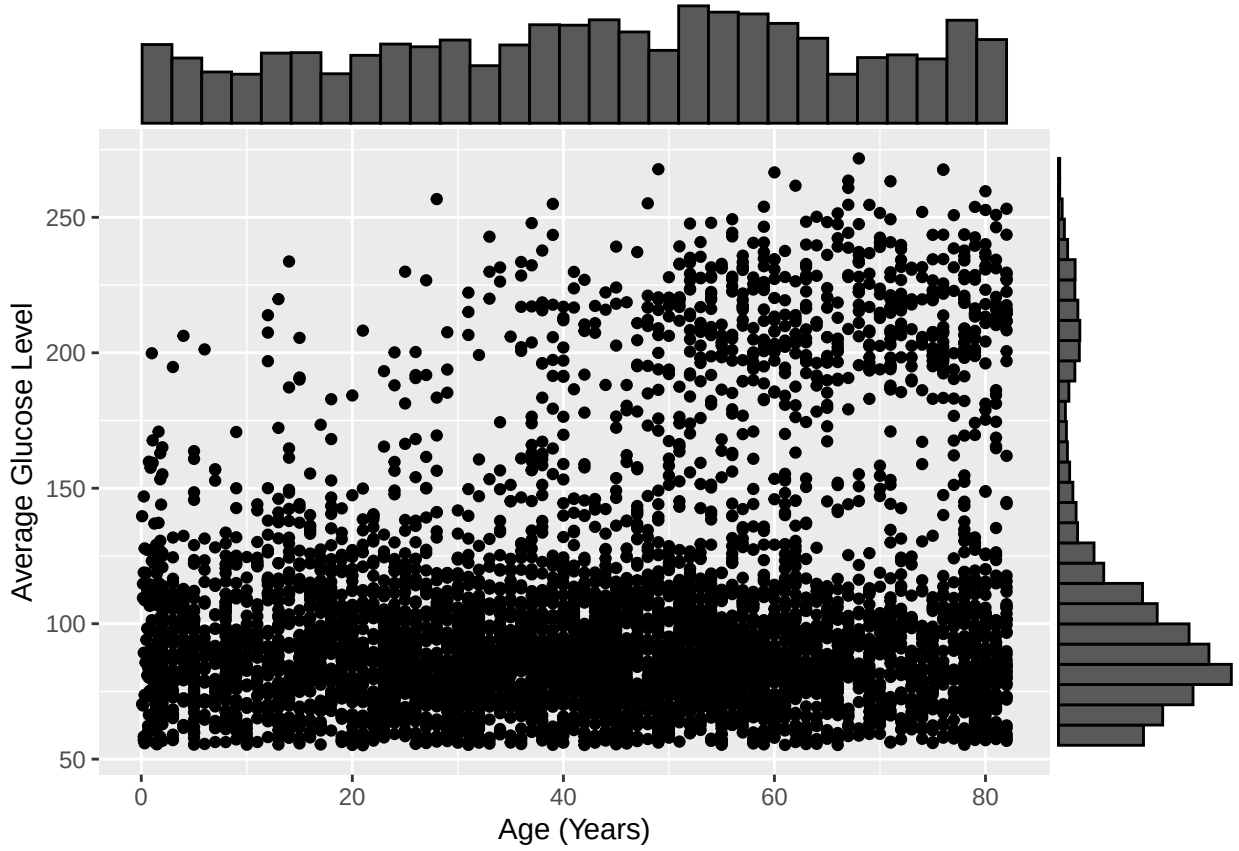


Graph from "More Data, More Problems" Slide

```
library(ggExtra)
```

```
## Warning: package 'ggExtra' was built under R version 4.0.5
```

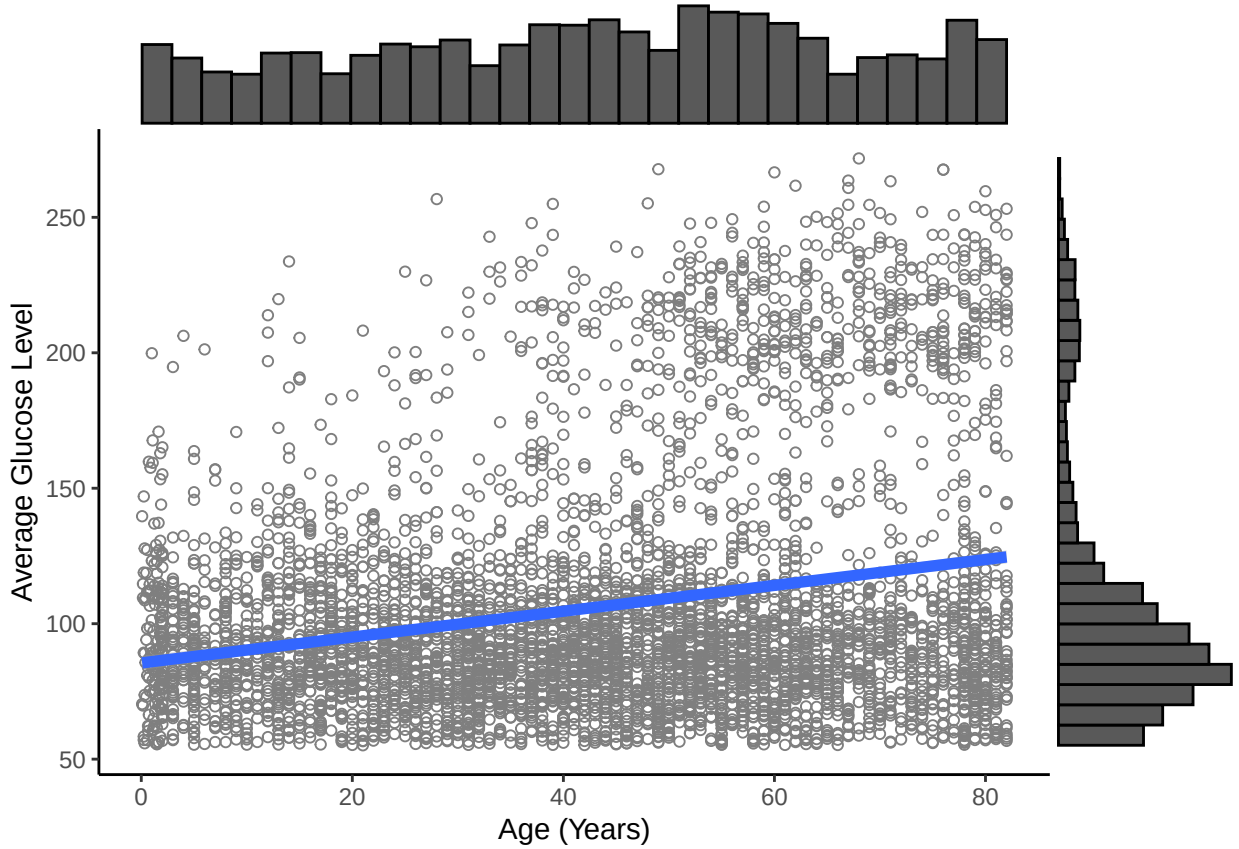
```
#Basic Scatter plot with histograms
a=ggplot(data,aes(x=age,y=avg_glucose_level))+
  geom_point()+
  xlab("Age (Years)")+
  ylab("Average Glucose Level")

ggMarginal(a,type = "histogram")
```



Graph from "Visualize Linear Trend" Slide

```
library(ggExtra)
#Basic Scatter plot with histograms, and linear trend
b=ggplot(data,aes(x=age,y=avg_glucose_level))+
  geom_point(shape=21,color="gray50")+
  geom_smooth(method="lm",linewidth=2)+
  xlab("Age (Years)")+
  ylab("Average Glucose Level")+
  theme_classic()


ggMarginal(b,type = "histogram")
```

```
##

## `geom_smooth()` using formula = 'y ~ x'
##
##
## `geom_smooth()` using formula = 'y ~ x'
```



Graph from "Group Trends" Slide

Here we will fit the linear trends for people with and without stroke as a function of age (independent variable) and average glucose (dependent variable).

To do so we must first fit a linear model and then using the resulting model to generate the predicted line across the entire age range of our data.

```
#Define who does and does not have a stroke (stroke_status)
data$stroke_status=ifelse(data$stroke==1,"Stroke","No Stroke")

#Fit the linear model of age and stroke status to average glucose level.
fit=lm(avg_glucose_level~stroke_status+age,data)
summary(fit)
```

```
##
## Call:
## lm(formula = avg_glucose_level ~ stroke_status + age, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -80.94 -28.92 -12.22  13.41 159.89
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         86.36795    1.33265  64.809  < 2e-16 ***
## stroke_statusStroke 16.45592    2.93922   5.599 2.27e-08 ***
```

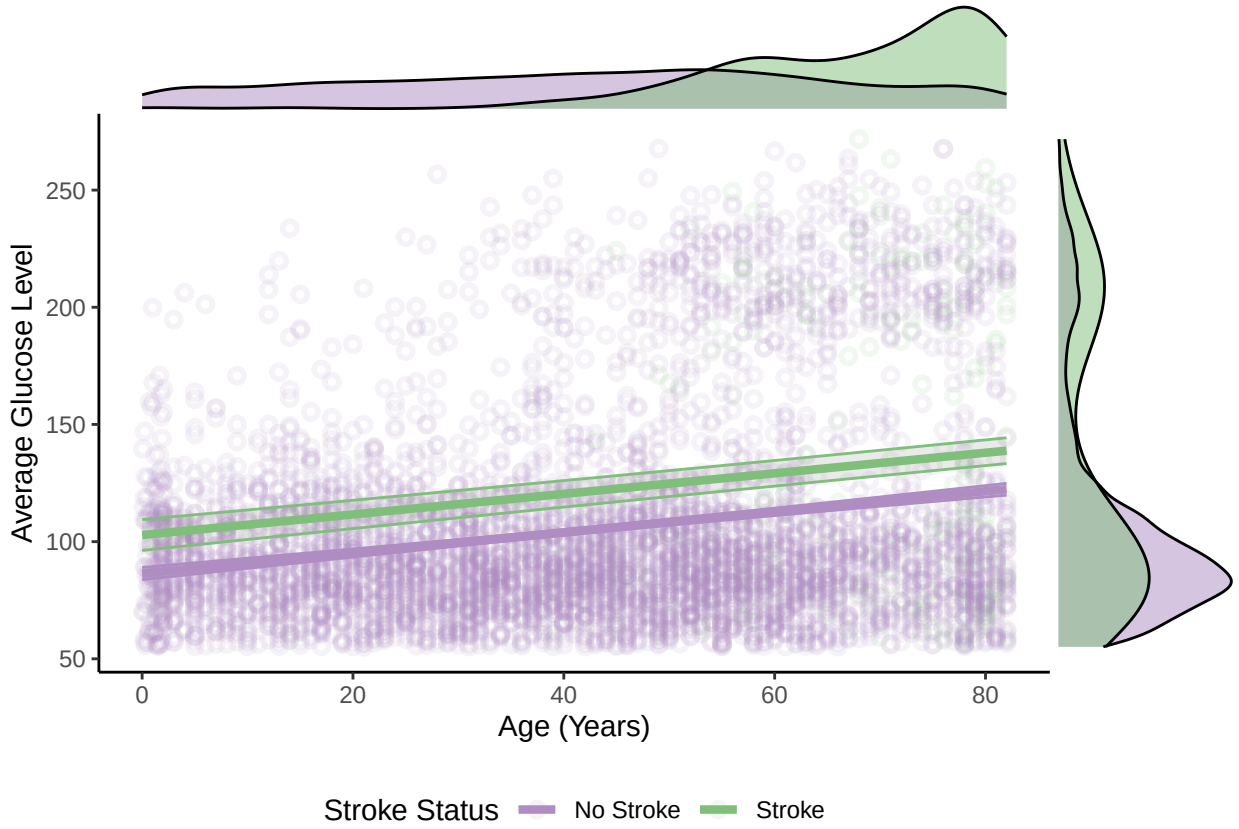
```
## age                  0.43883    0.02799  15.679  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.85 on 5106 degrees of freedom
## Multiple R-squared:  0.06255,    Adjusted R-squared:  0.06219
## F-statistic: 170.4 on 2 and 5106 DF,  p-value: < 2.2e-16
```

```
#create a matrix of "new data" that will generate the trend across the entire
#age range for people with stroke.
new.data=expand.grid(age=c(0:82),
                     stroke_status=c("Stroke"))

#Generate the standard error for the fitted line using the fitted model (fit)
se.fit.s=predict(fit,new.data,se.fit = T)

#A little math to generate the confidence interval across the trend
new.data$lwr=se.fit.s$fit-(1.96*se.fit.s$se.fit) #lower level
new.data$pred=se.fit.s$fit #predict glucose
new.data$upr=se.fit.s$fit+(1.96*se.fit.s$se.fit) #upper level

#Now let's do the same for people without Stroke.
new.data.no=expand.grid(age=c(0:82),
                     stroke_status=c("No Stroke"))

se.fit=predict(fit,new.data.no,se.fit = T)

new.data.no$lwr=se.fit$fit-(1.96*se.fit$se.fit)
new.data.no$pred=se.fit$fit
new.data.no$upr=se.fit$fit+(1.96*se.fit$se.fit)

#We will use these fitted lines for plotting.

c=ggplot(data,aes(x=age,y=avg_glucose_level,color=stroke_status))+
  geom_point(shape=21,stroke=1.5,alpha=0.1)+
  # geom_point(data=data[data$stroke_status=="Stroke",],mapping=aes(x=age,y=avg_glucose_level),
  #            color="#7fbf7b",shape=21,stroke=1.5,alpha=.75)+
  scale_color_manual(values = c("#af8dc3","#7fbf7b"))+
  geom_ribbon(new.data,mapping=aes(x=age,ymin = lwr,ymax=upr),
              alpha=.1,color="#7fbf7b",inherit.aes = F)+
  geom_line(new.data,mapping=aes(x=age,y=pred),size=1.5)+
  geom_ribbon(new.data.no,mapping=aes(x=age,ymin = lwr,ymax=upr),
              alpha=.1,color="#af8dc3",inherit.aes = F)+
  geom_line(new.data.no,mapping=aes(x=age,y=pred),size=1.5)+
  xlab("Age (Years)")+
  ylab("Average Glucose Level")+
  labs(color="Stroke Status")+
  theme_classic()+
  theme(legend.position = "bottom")
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
```

```
## generated.
```

```
ggMarginal(c,type = "density",groupFill = T)
```




```
#Pretty good but it's difficult to see the individual data.
```

Graph from "Group Trends" Slide 2

```
d=ggplot(data,aes(x=age,y=avg_glucose_level,color=stroke_status))+
  geom_point(shape=21,stroke=1.5,alpha=0.1)+
  #Increase the boldness of the stroke data with the aesthetic "stroke"
  geom_point(data=data[data$stroke_status=="Stroke",],mapping=aes(x=age,y=avg_glucose_level),
             color="#7fbf7b",shape=21,stroke=1.5,alpha=.75)+
  scale_color_manual(values = c("#af8dc3","#7fbf7b"))+
  geom_ribbon(new.data,mapping=aes(x=age,ymin = lwr,ymax=upr),
              alpha=.1,color="#7fbf7b",inherit.aes = F)+
  geom_line(new.data,mapping=aes(x=age,y=pred),size=1.5)+
  geom_ribbon(new.data.no,mapping=aes(x=age,ymin = lwr,ymax=upr),
              alpha=.1,color="#af8dc3",inherit.aes = F)+
  geom_line(new.data.no,mapping=aes(x=age,y=pred),size=1.5)+
  xlab("Age (Years)")+
  ylab("Average Glucose Level")+
  labs(color="Stroke Status")+
```

```
  theme_classic()+
  theme(legend.position = "bottom")

ggMarginal(d,type = "density",groupFill = T)
```