

# Data Visualization

## Contents

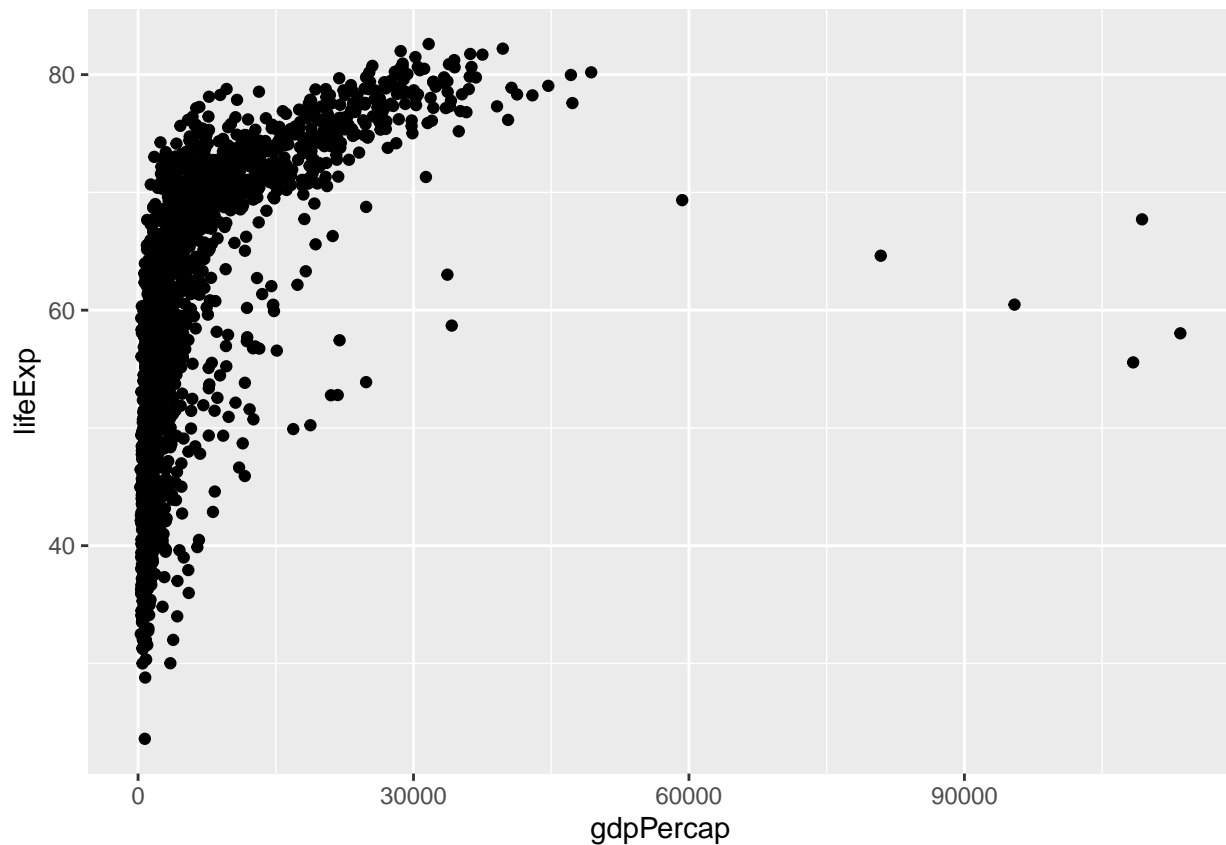
<b>1</b>	<b>Scatter Plots</b>	<b>2</b>
<b>2</b>	<b>Lines Plots</b>	<b>13</b>
<b>3</b>	<b>Facets</b>	<b>18</b>
<b>4</b>	<b>Barplots</b>	<b>22</b>
<b>5</b>	<b>Histograms</b>	<b>29</b>
<b>6</b>	<b>Boxplots</b>	<b>38</b>
<b>7</b>	<b>Dotted Plots</b>	<b>40</b>
<b>8</b>	<b>Heatmap</b>	<b>42</b>
<b>9</b>	<b>Violin Plots</b>	<b>43</b>
<b>10</b>	<b>References</b>	<b>45</b>

```
library("ggplot2")
library("gapminder")
library("dplyr")
glimpse(gapminder)

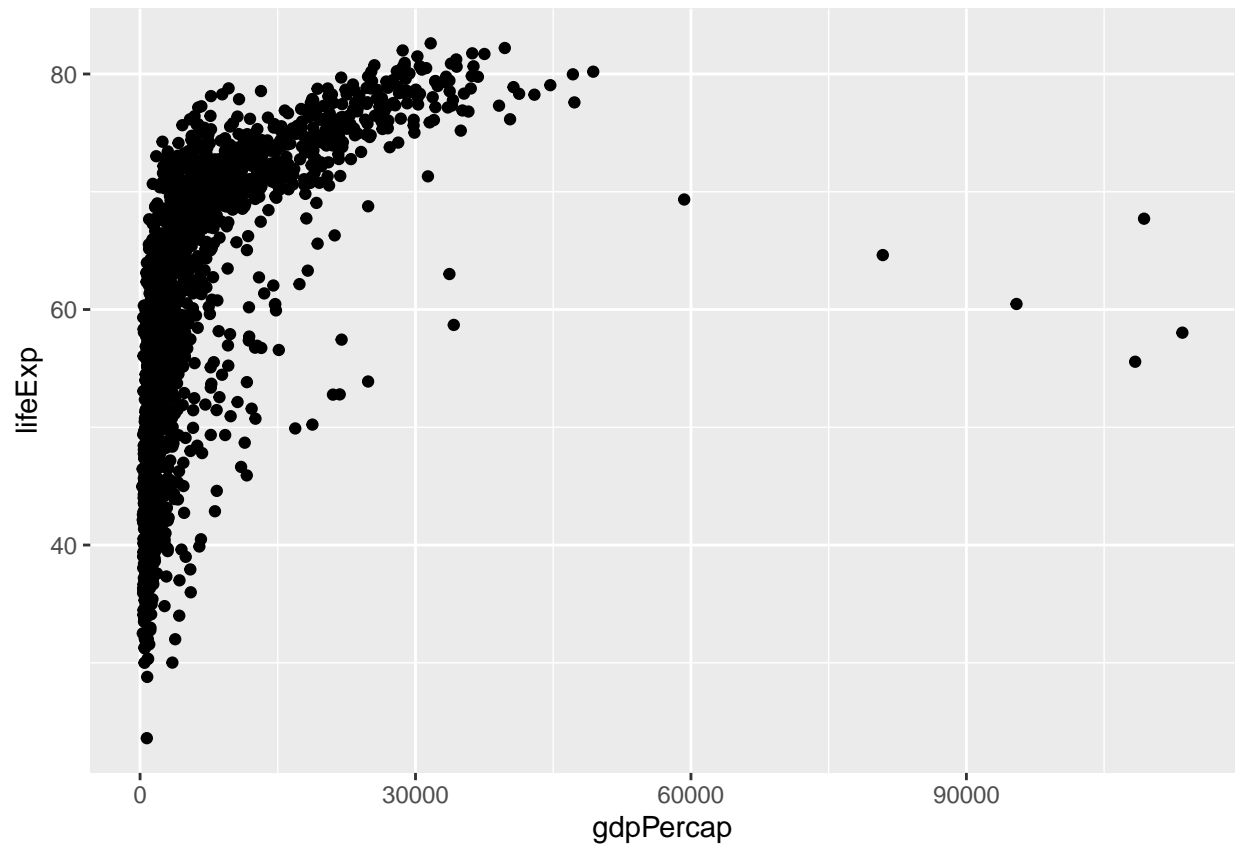
## Rows: 1,704
## Columns: 6
## $ country   <fct> Afghanistan, Afghanistan, Afghanistan, Afghanistan, Afgha...
## $ continent <fct> Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asi...
## $ year      <int> 1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992, 199...
## $ lifeExp   <dbl> 28.801, 30.332, 31.997, 34.020, 36.088, 38.438, 39.854, 4...
## $ pop       <int> 8425333, 9240934, 10267083, 11537966, 13079460, 14880372,...
## $ gdpPercap <dbl> 779.4453, 820.8530, 853.1007, 836.1971, 739.9811, 786.113...
```

## 1 Scatter Plots

```
# basic scatterplots
ggplot(data = gapminder,
       mapping = aes(x = gdpPercap, y = lifeExp)) + geom_point()
```

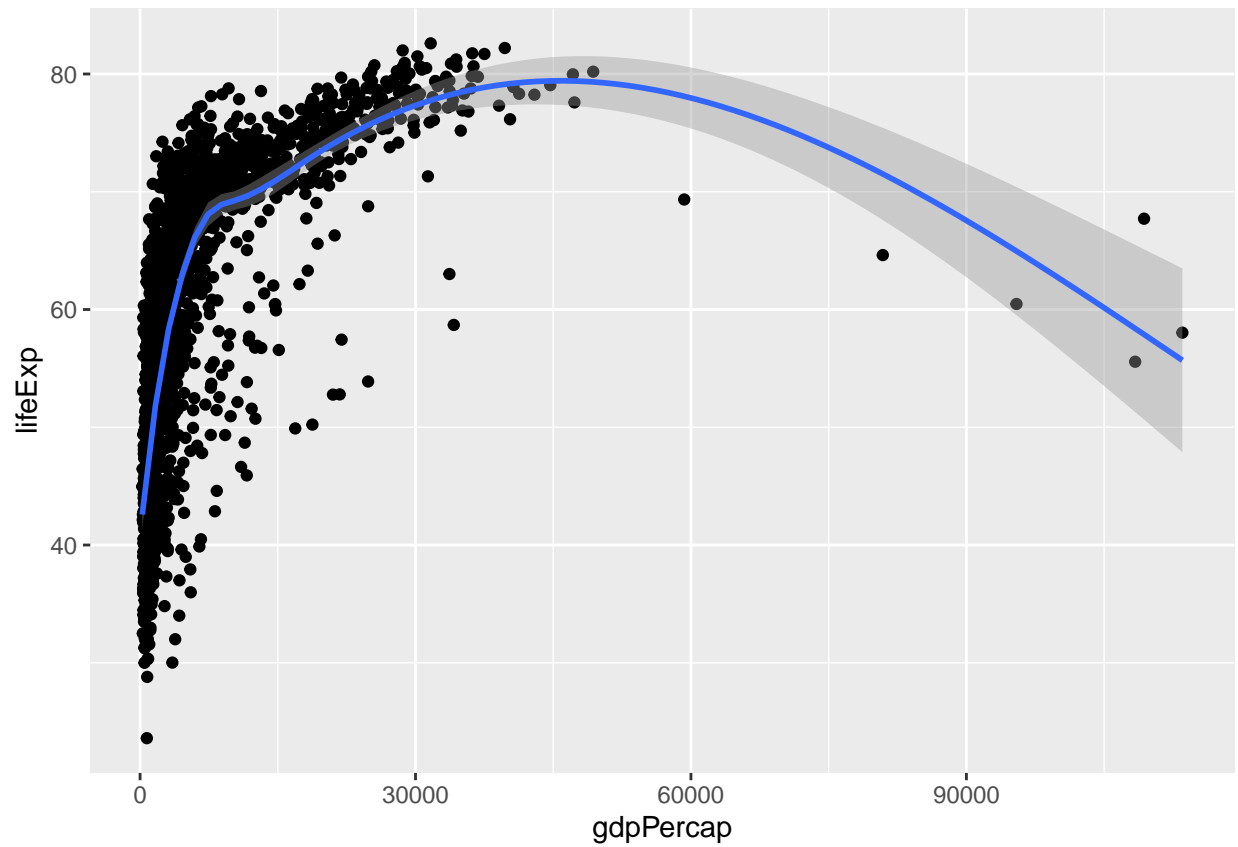


```
ggplot(gapminder, aes(x = gdpPercap, y = lifeExp)) + geom_point()
```



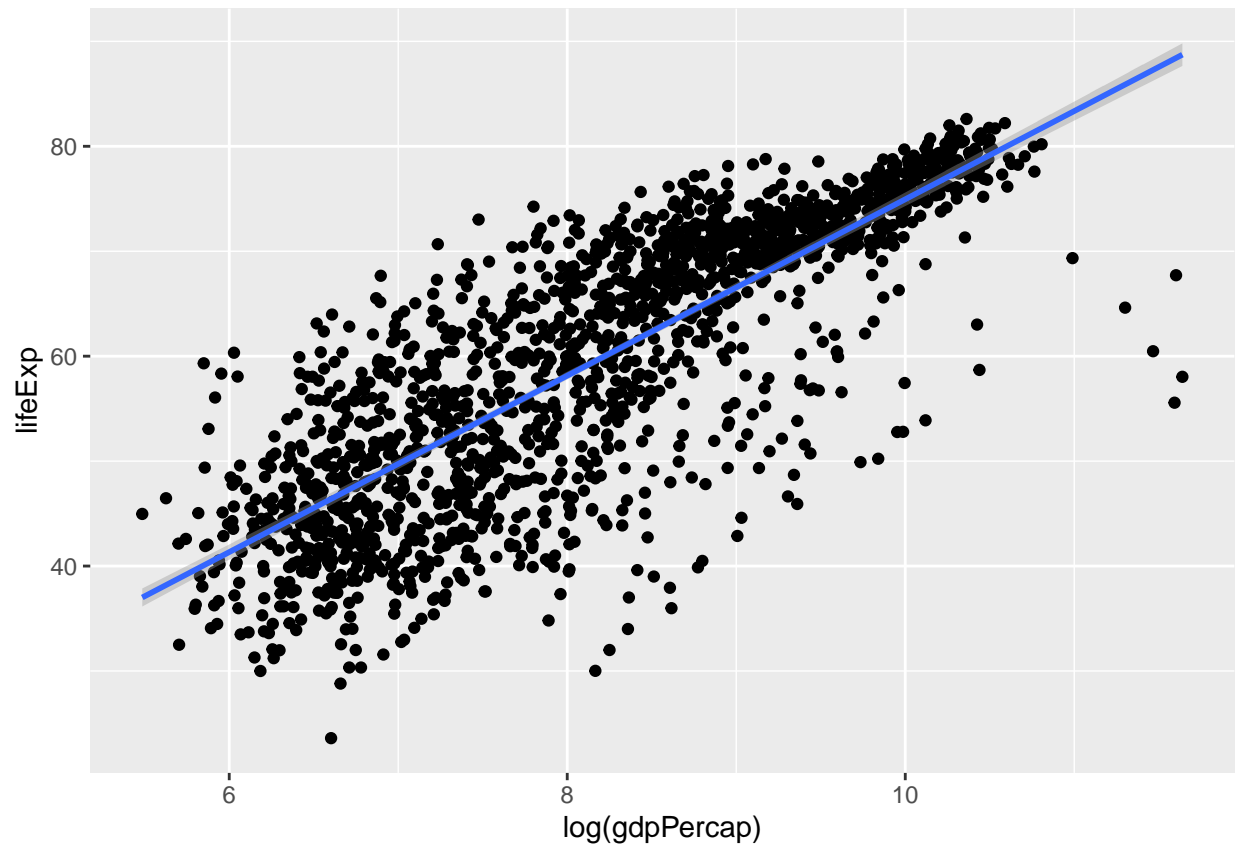
```
# adding smoothing lines
ggplot(gapminder,aes(x = gdpPercap,y = lifeExp)) +
  geom_point() +
  geom_smooth()

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



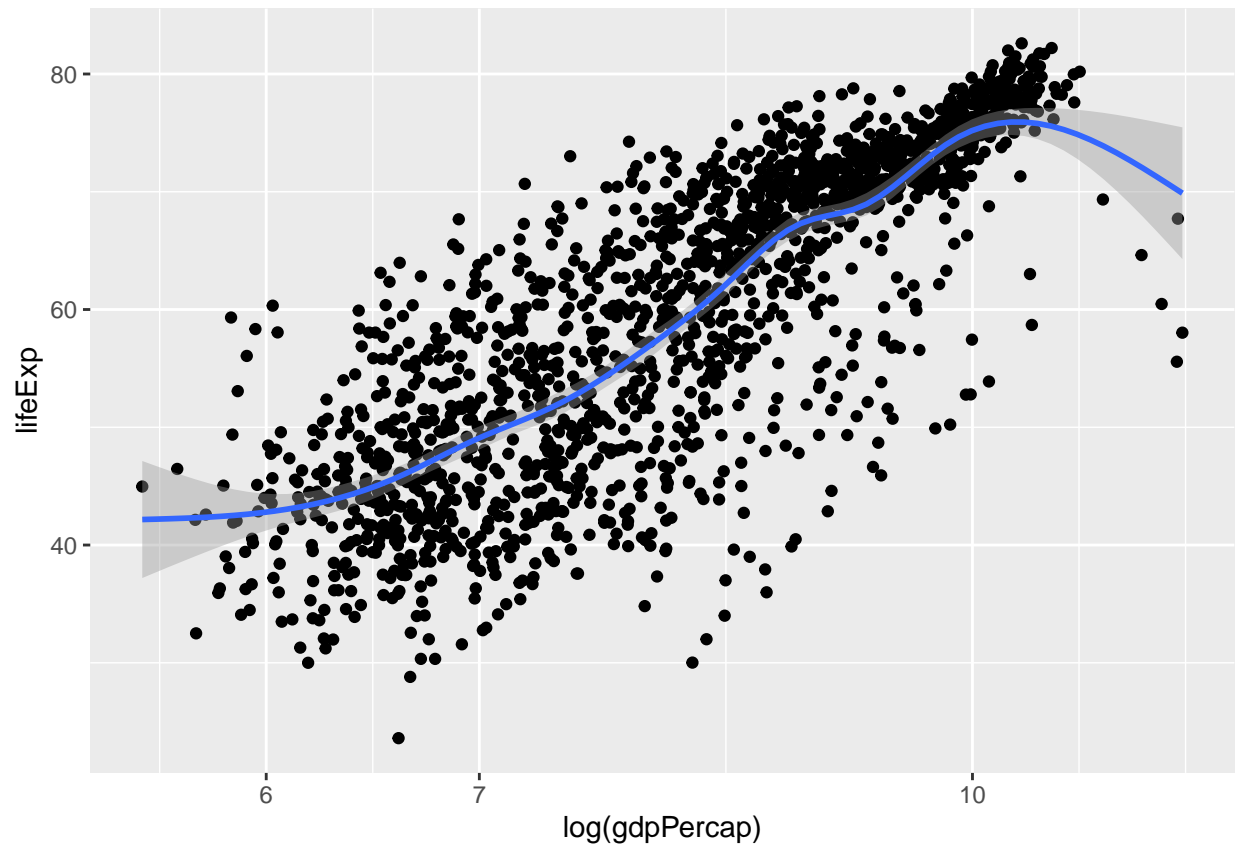
```
ggplot(gapminder, aes(x = log(gdpPerCap), y = lifeExp)) +  
  geom_point() +  
  geom_smooth(method = "lm")
```

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



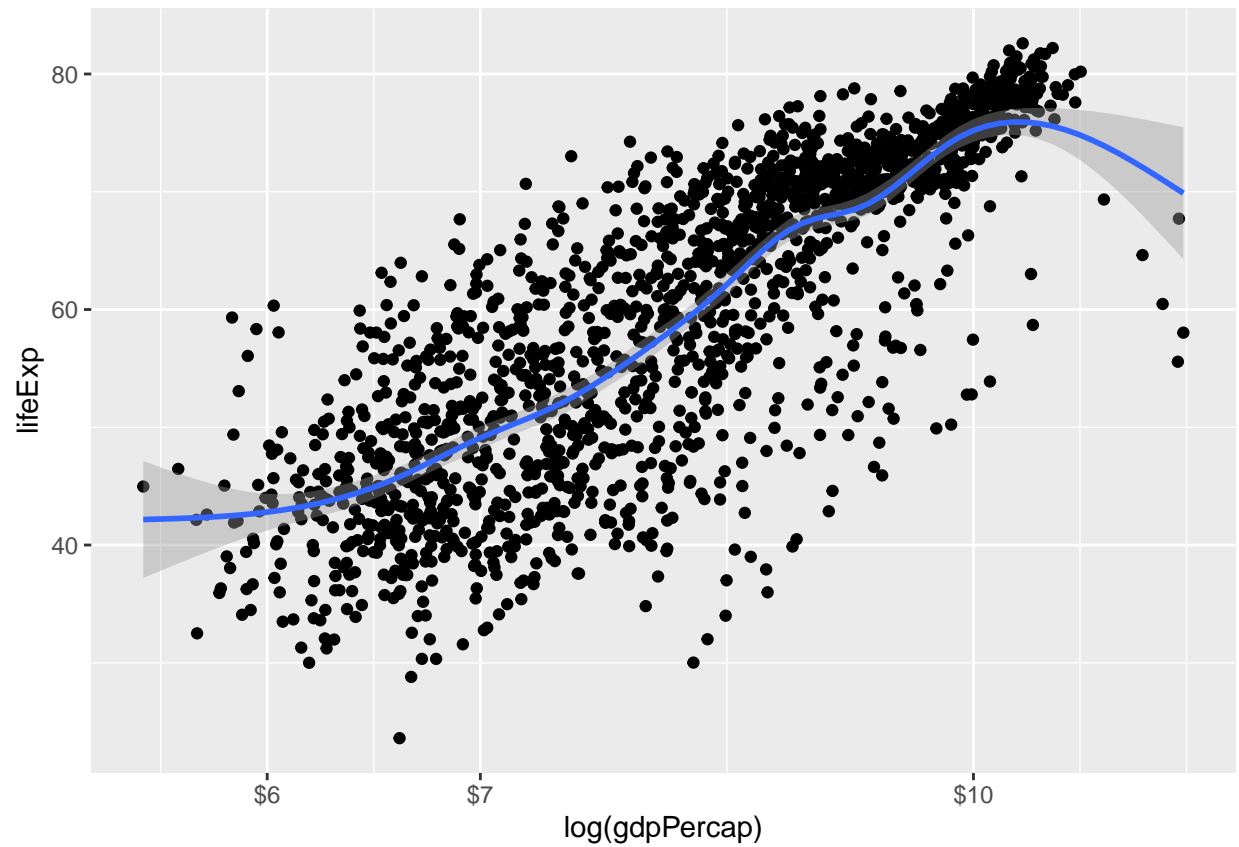
```
ggplot(gapminder, aes(x = log(gdpPerCap), y = lifeExp)) +  
  geom_point() +  
  geom_smooth(method="gam") + #generalized additive model  
  scale_x_log10()
```

```
## `geom_smooth()` using formula 'y ~ s(x, bs = "cs")'
```

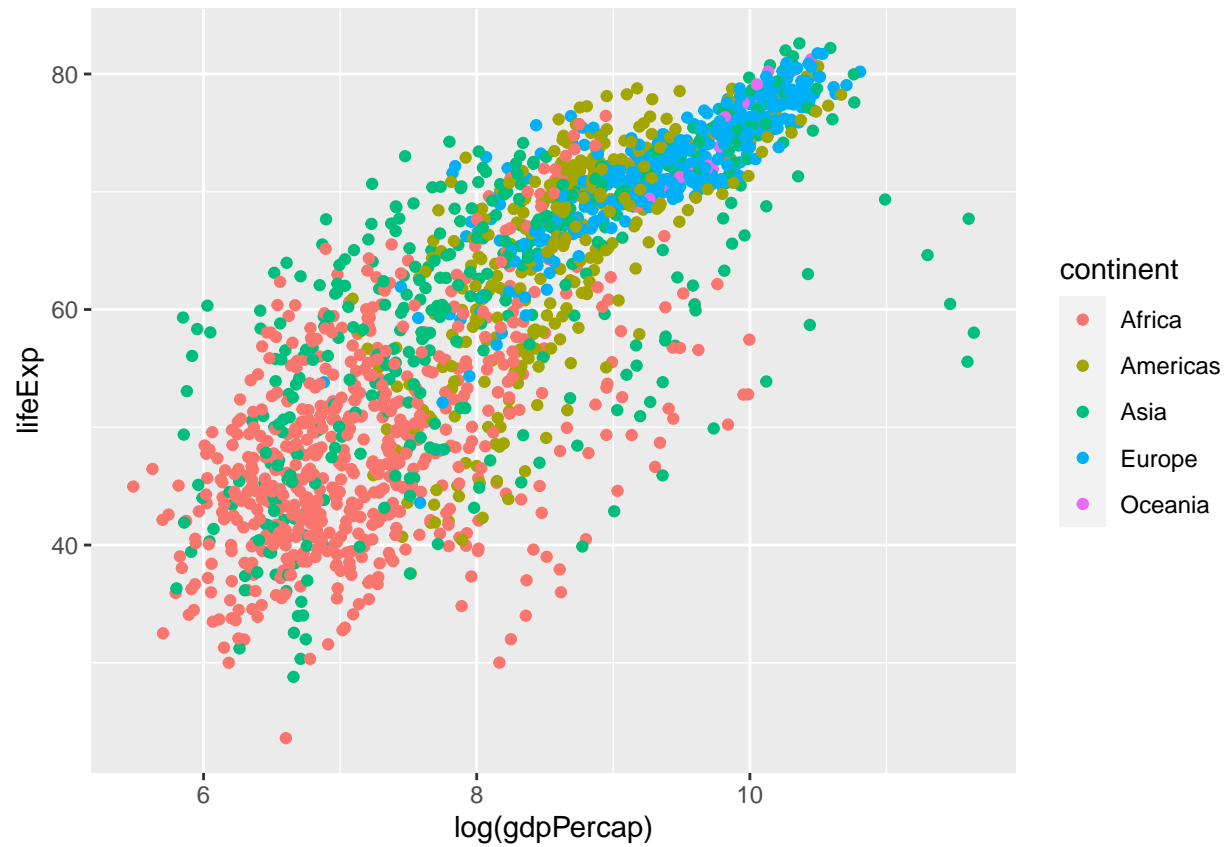


```
ggplot(gapminder, aes(x = log(gdpPerCap), y = lifeExp)) +  
  geom_point() +  
  geom_smooth(method="gam") +  
  scale_x_log10(labels=scales::dollar)
```

```
## `geom_smooth()` using formula 'y ~ s(x, bs = "cs")'
```



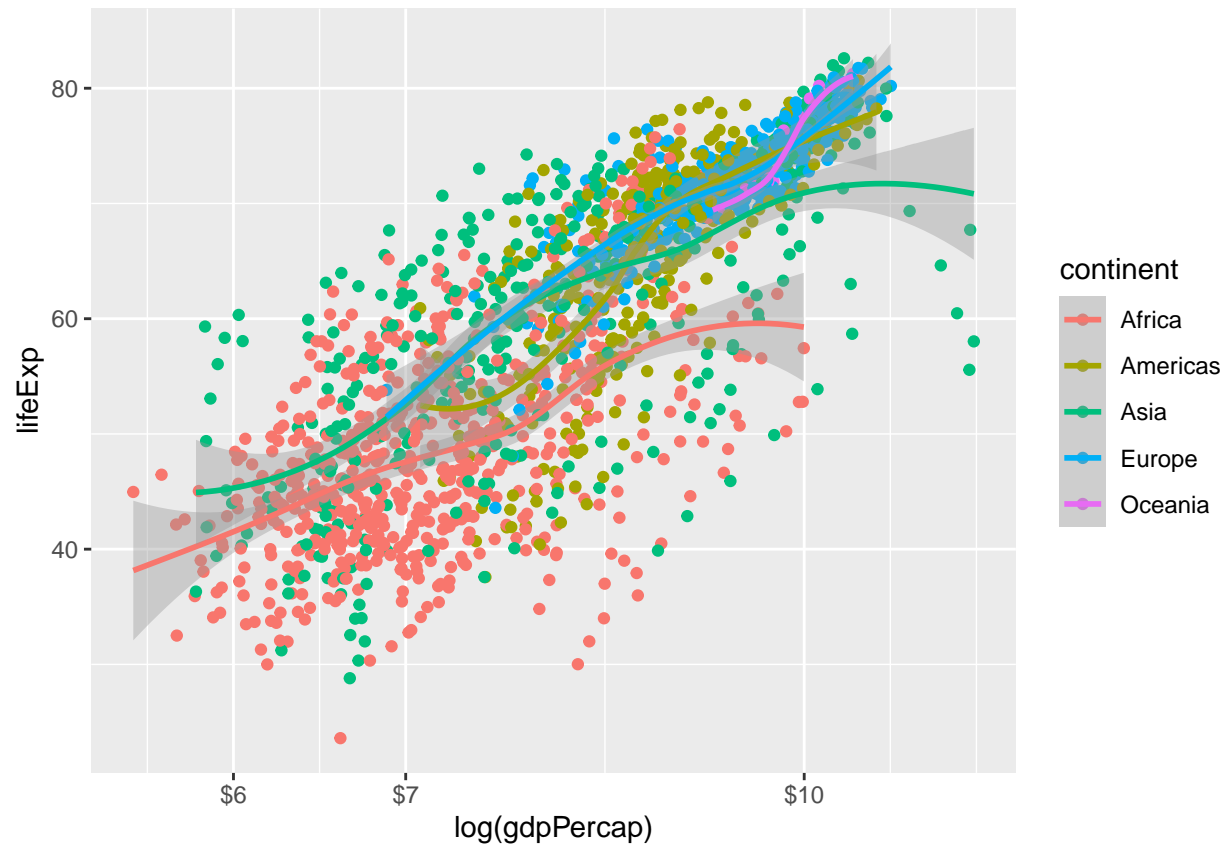
```
# controlling colors, shapes, linetype, and transparency of points  
ggplot(gapminder, aes(x = log(gdpPercap), y = lifeExp, color = continent)) +  
  geom_point()
```



```
ggplot(gapminder, aes(x = log(gdpPercap), y = lifeExp, color = continent)) +  
  geom_point() +  
  geom_smooth(method="loess") +  
  scale_x_log10(labels=scales::dollar)
```

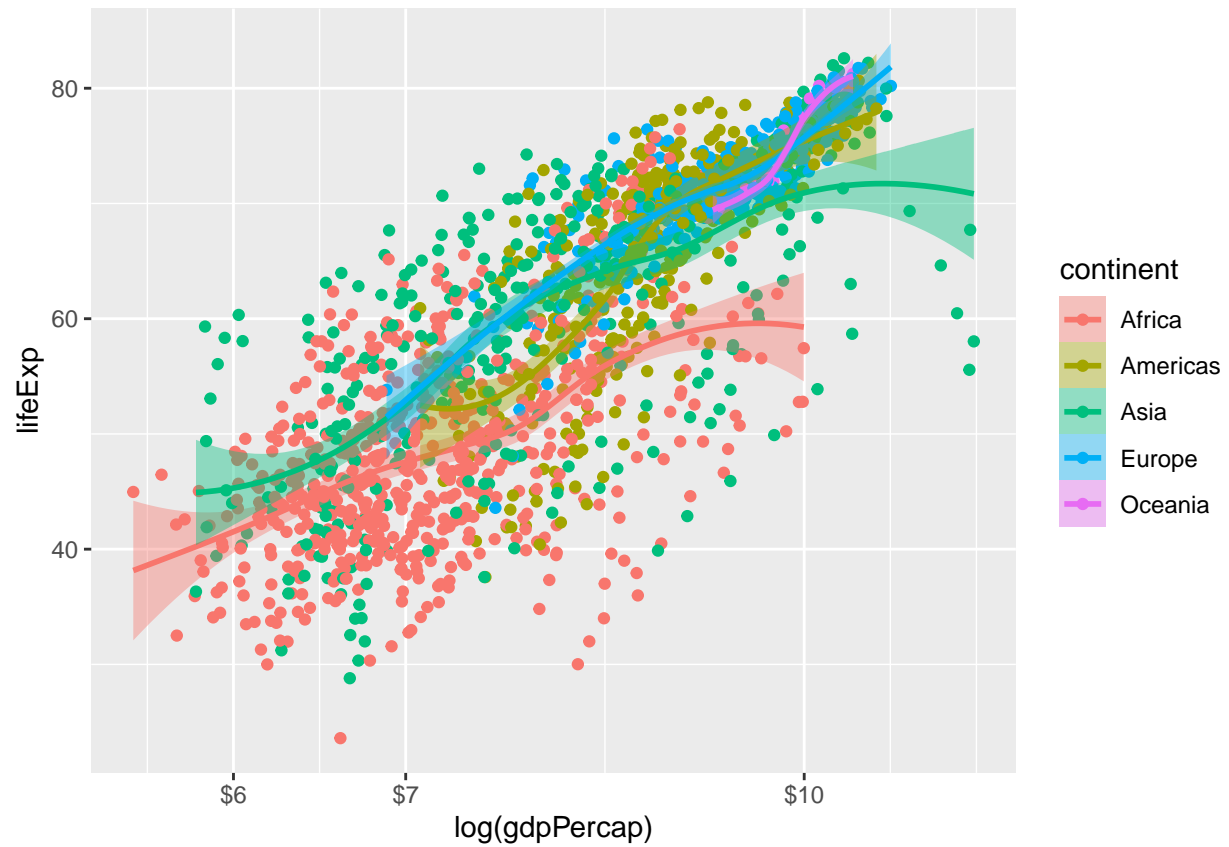
```
## `geom_smooth()` using formula 'y ~ x'
```





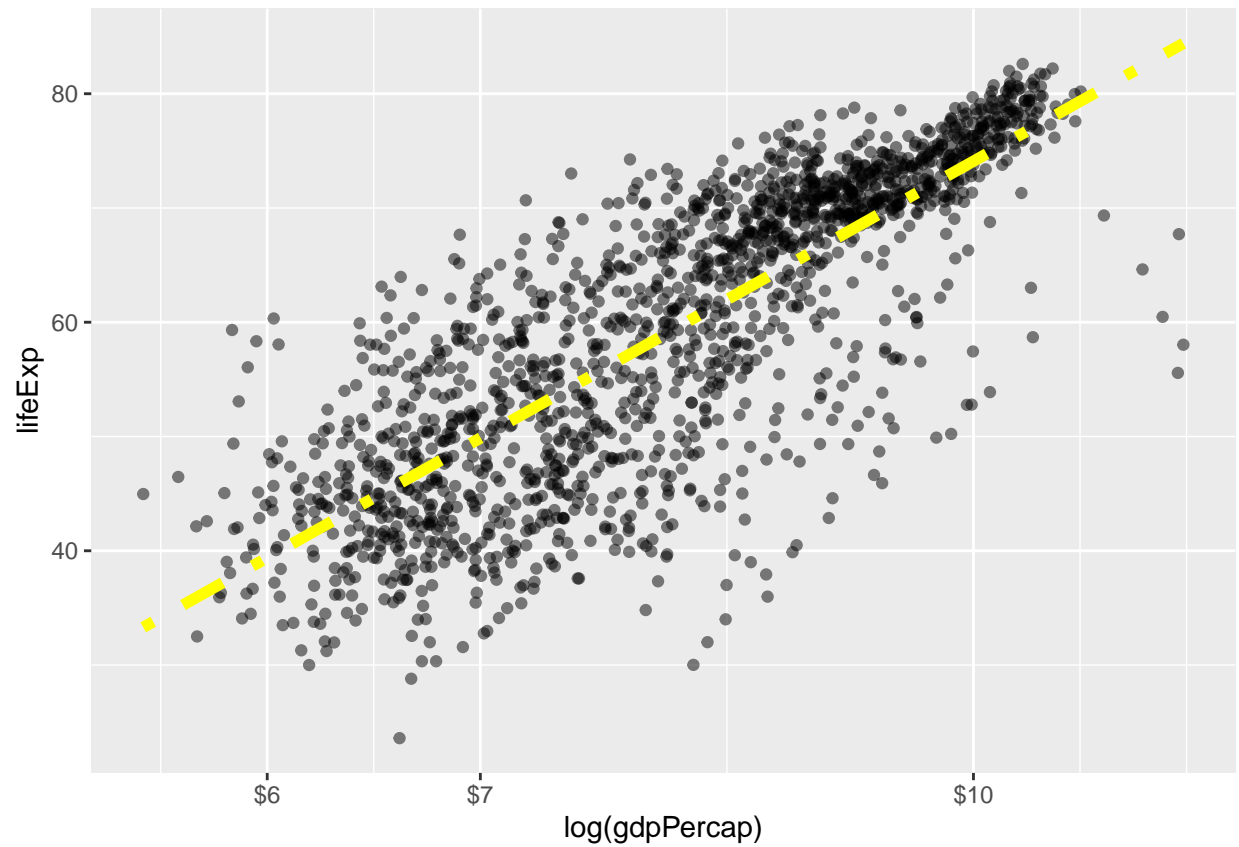
```
ggplot(gapminder, aes(x = log(gdpPercap), y = lifeExp, color = continent, fill = continent)) +
  geom_point() +
  geom_smooth(method="loess") +
  scale_x_log10(labels=scales::dollar)
```

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

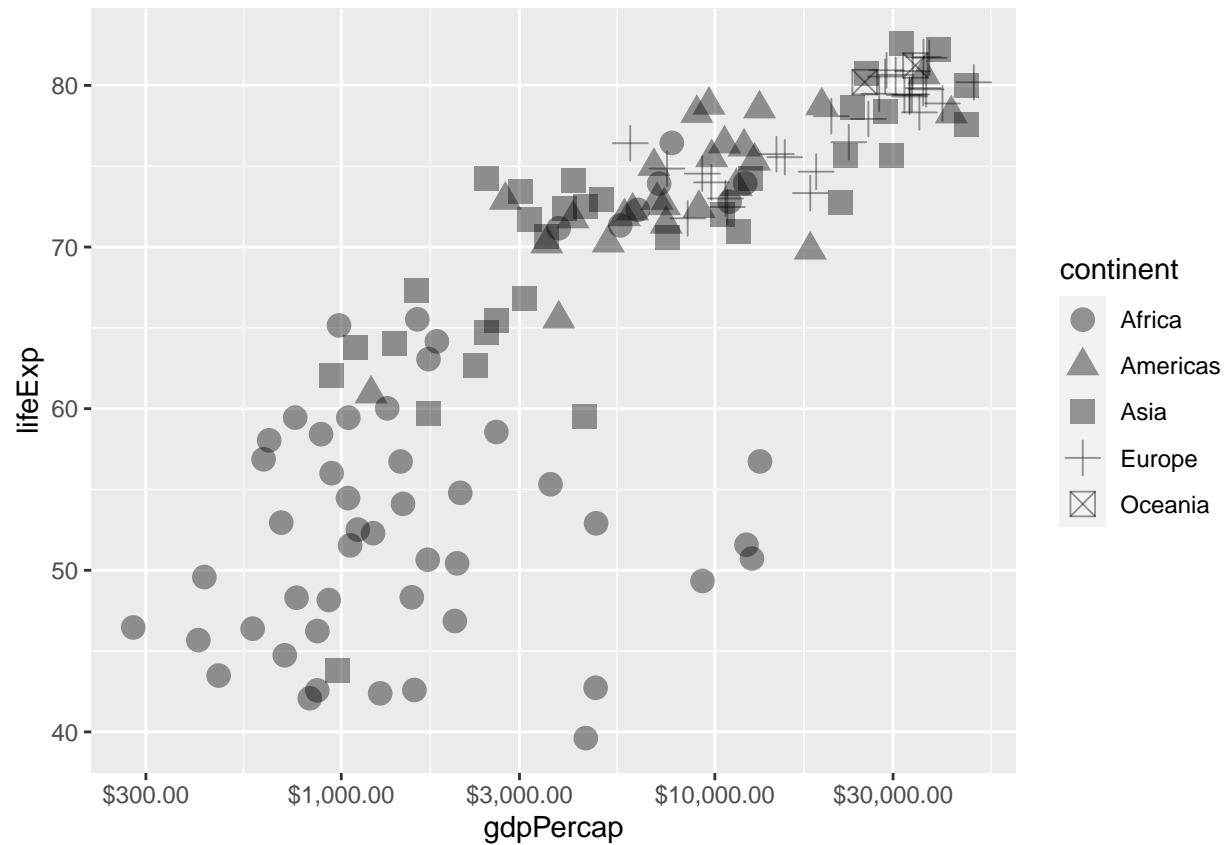


```
ggplot(gapminder, aes(x = log(gdpPercap), y = lifeExp)) +
  geom_point(alpha=0.5) +
  geom_smooth(method="lm", color="yellow", se = FALSE, size = 2, alpha = 0.3, linetype = 4) +
  scale_x_log10(labels=scales::dollar)
```

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



```
ggplot(data = filter(gapminder, year == "2007"),  
  mapping = aes(  
    x = gdpPercap,  
    y = lifeExp,  
    shape = continent)) +  
  geom_point(alpha = 0.4, size = 4) +  
  scale_x_log10(labels=scales::dollar)
```



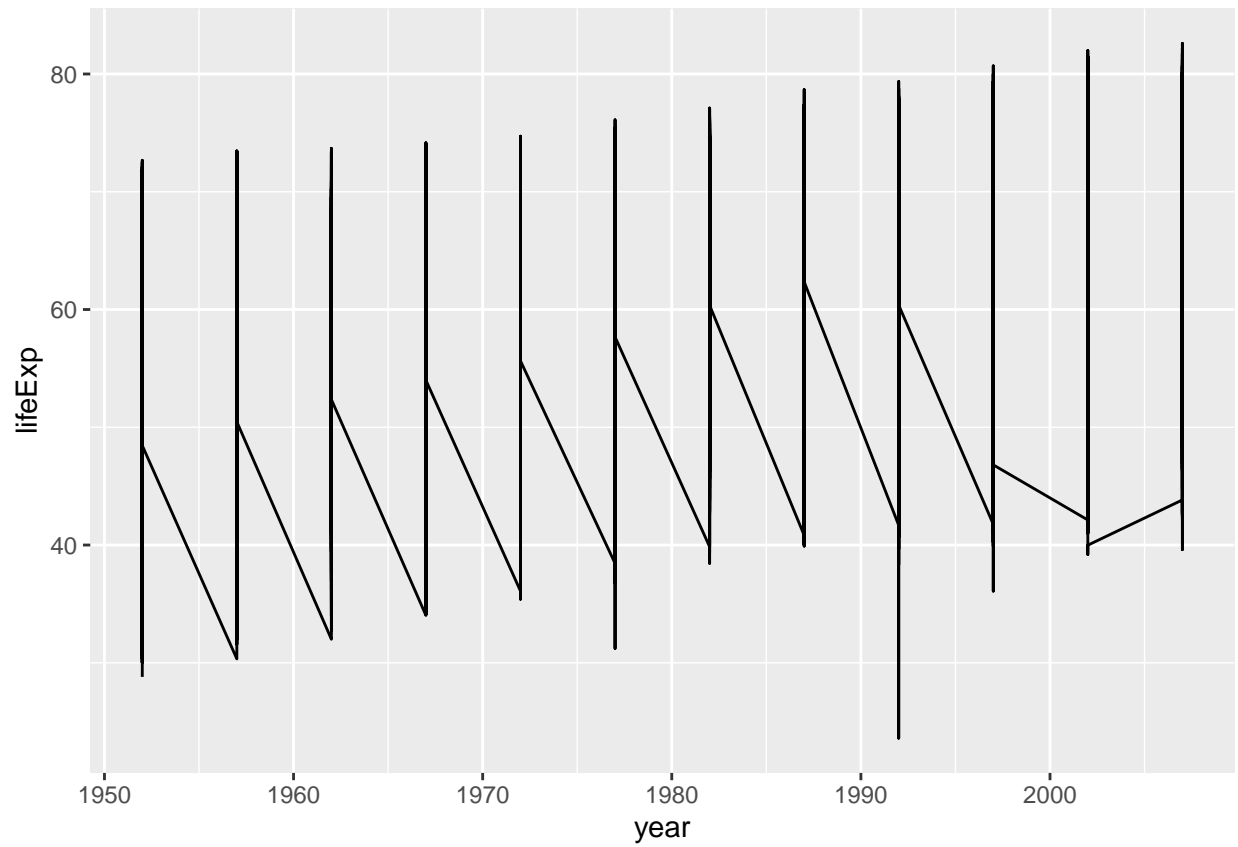
```
ggplot(data=gapminder,
       mapping = aes(
         x = gdpPercap,
         y = lifeExp,
         color = log(pop))) +
  geom_point() +
  geom_smooth(method="loess") +
  scale_x_log10(labels=scales::dollar)
```

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

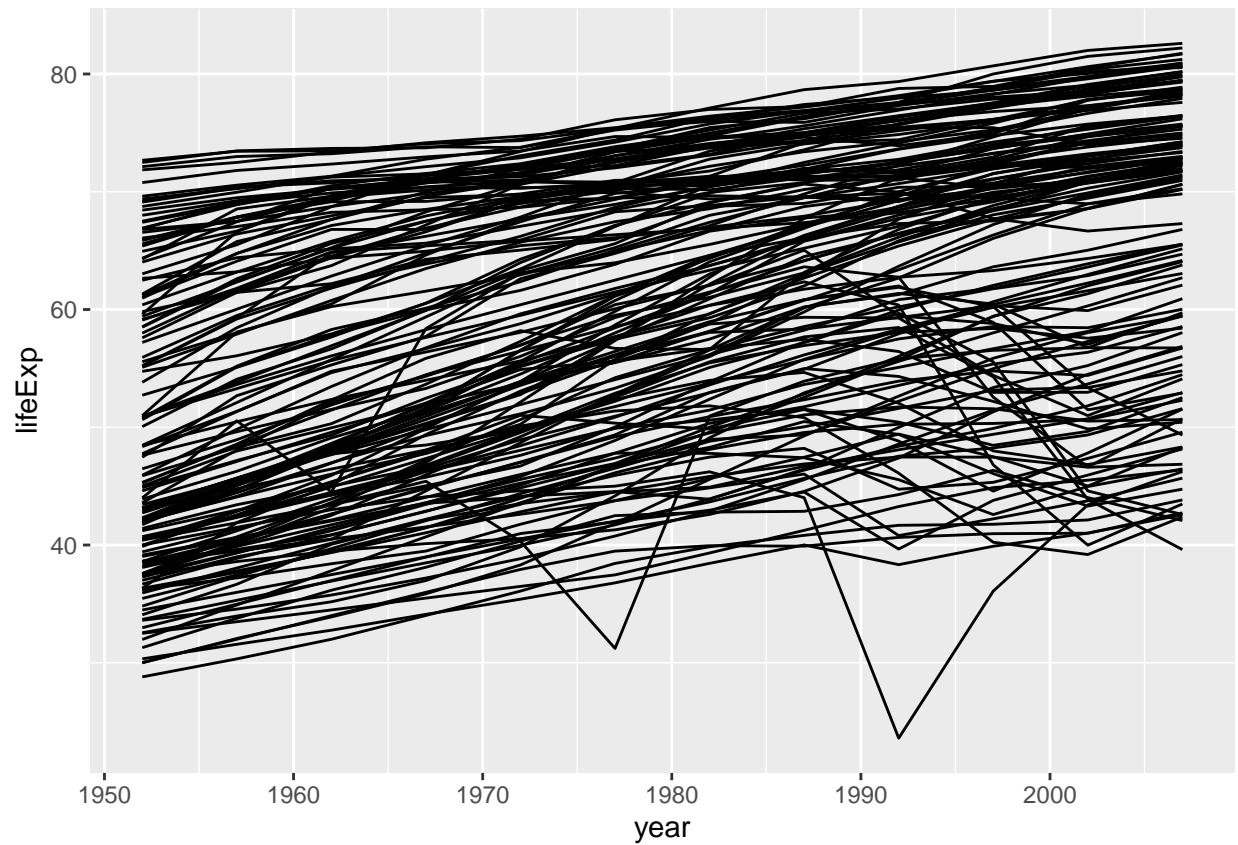


## 2 Lines Plots

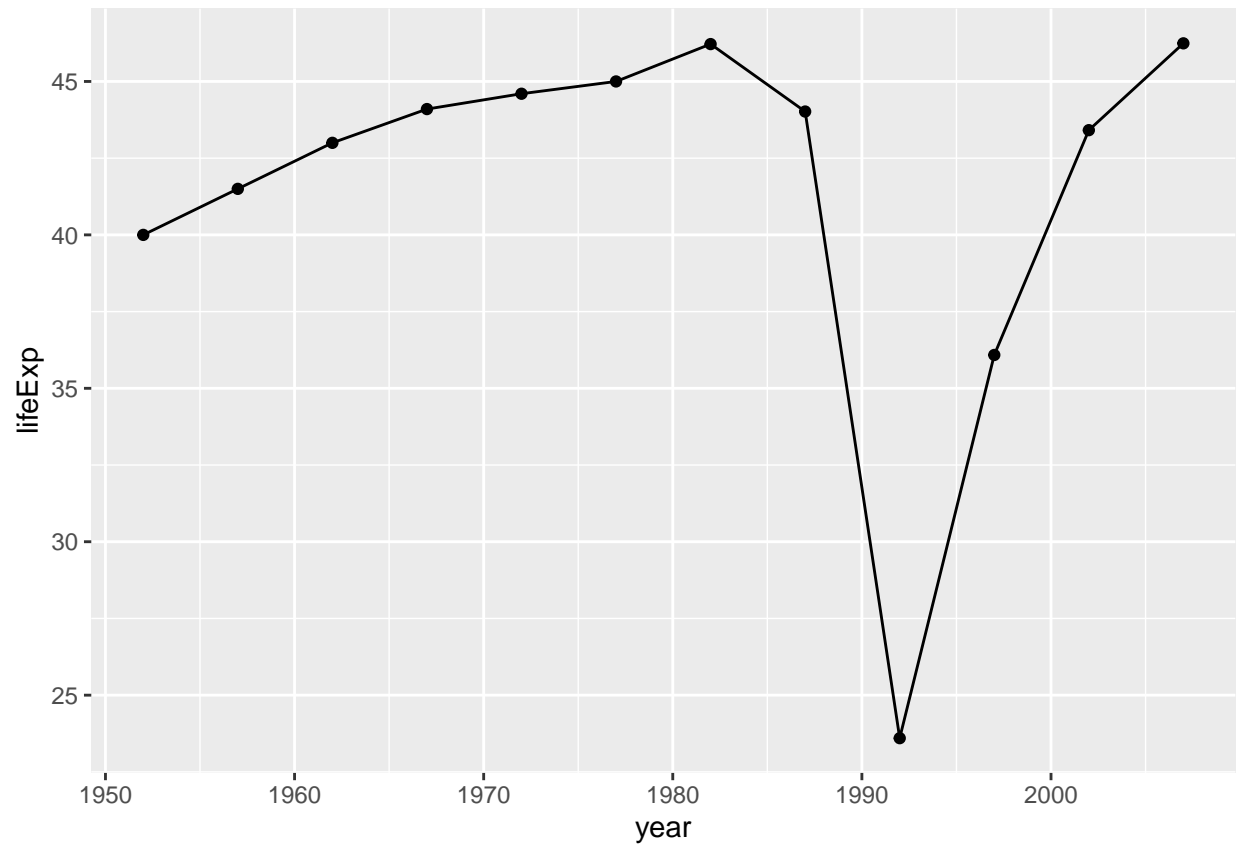
```
ggplot(data = gapminder,  
       mapping = aes(x = year, y = lifeExp)) +  
  geom_line()
```



```
ggplot(data = gapminder,  
       mapping = aes(x = year, y = lifeExp, group = country)) +  
  geom_line()
```

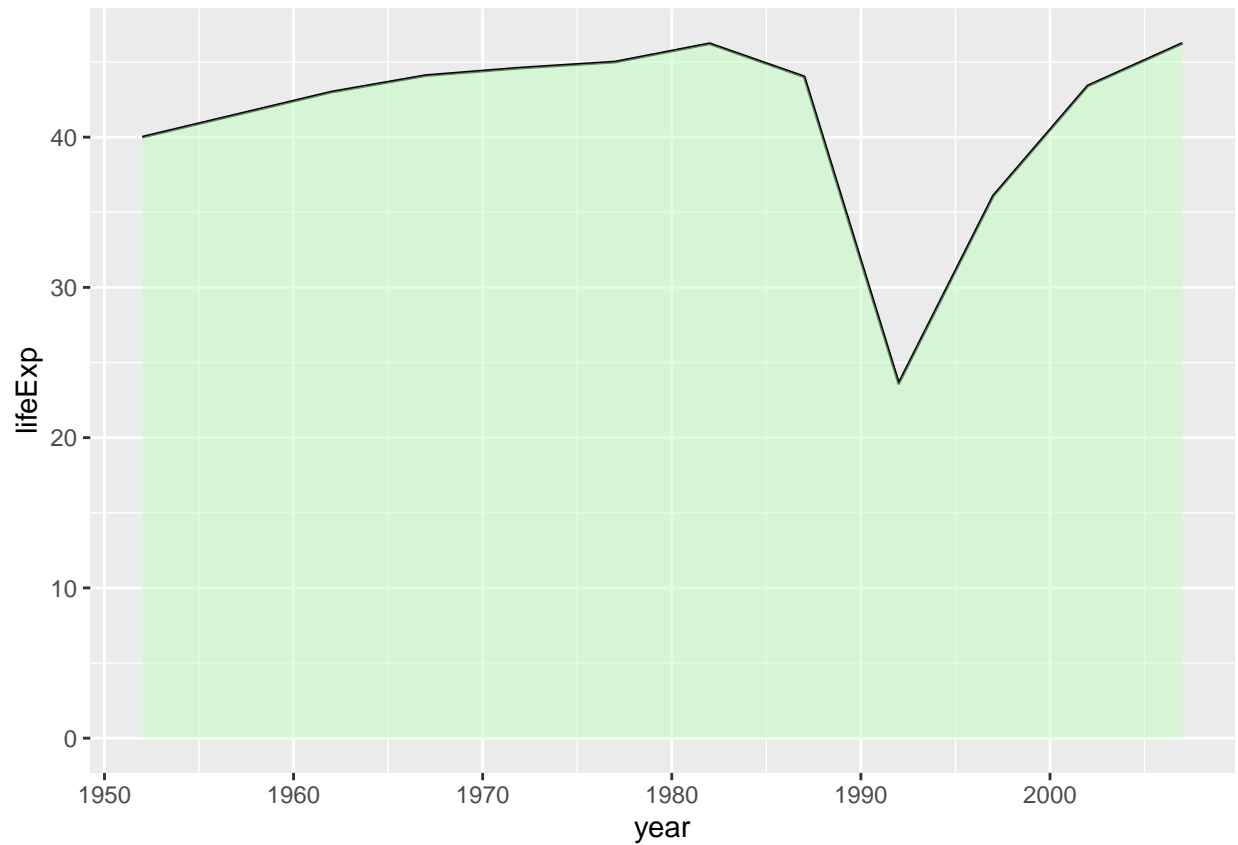


```
ggplot(data = filter(gapminder, country == "Rwanda"),  
       mapping = aes(x = year, y = lifeExp)) +  
  geom_line() +  
  geom_point()
```

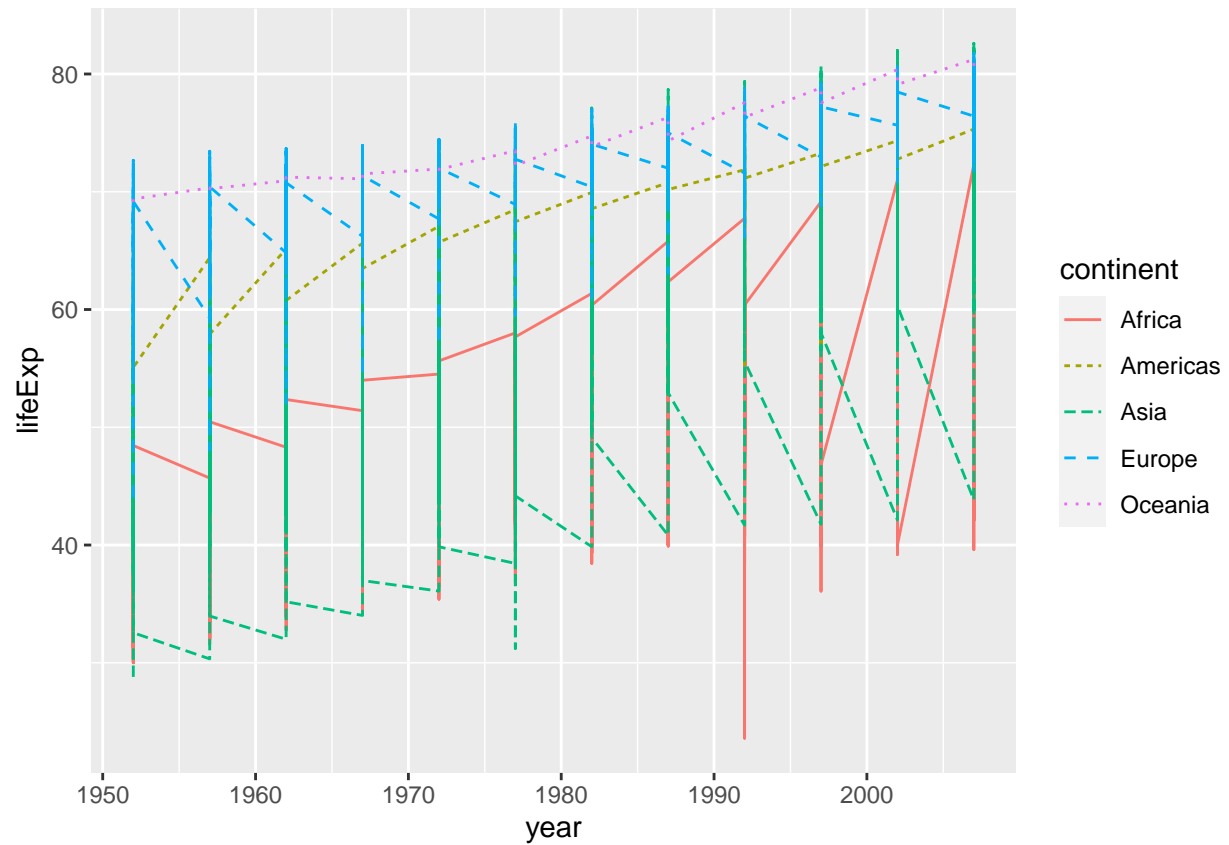


```
ggplot(data = filter(gapminder, country == "Rwanda"),  
       mapping = aes(x = year, y = lifeExp)) +  
  geom_line() +  
  geom_area(fill = "darkseagreen1", alpha = 0.5)
```



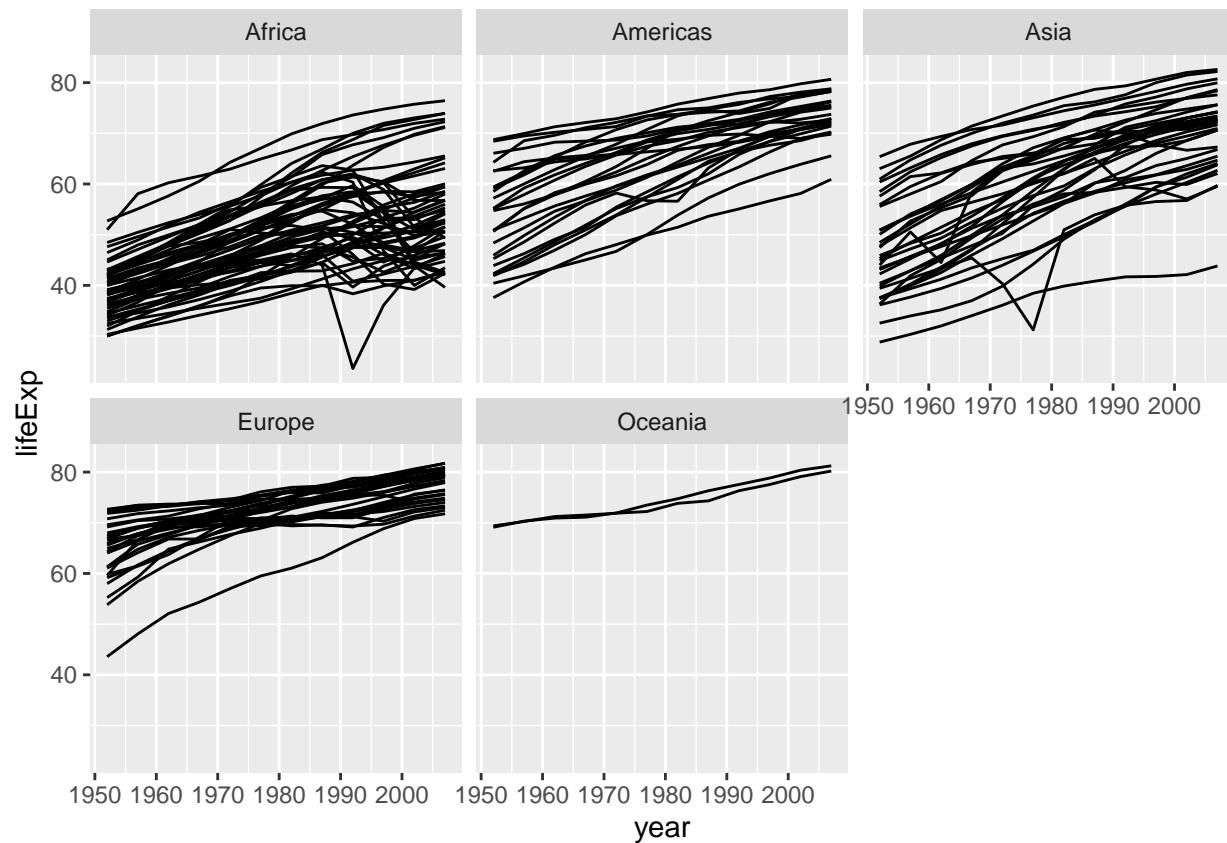


```
ggplot(data = gapminder,  
       mapping = aes(x = year, y = lifeExp)) +  
  geom_line(aes(color = continent, linetype = continent))
```



### 3 Facets

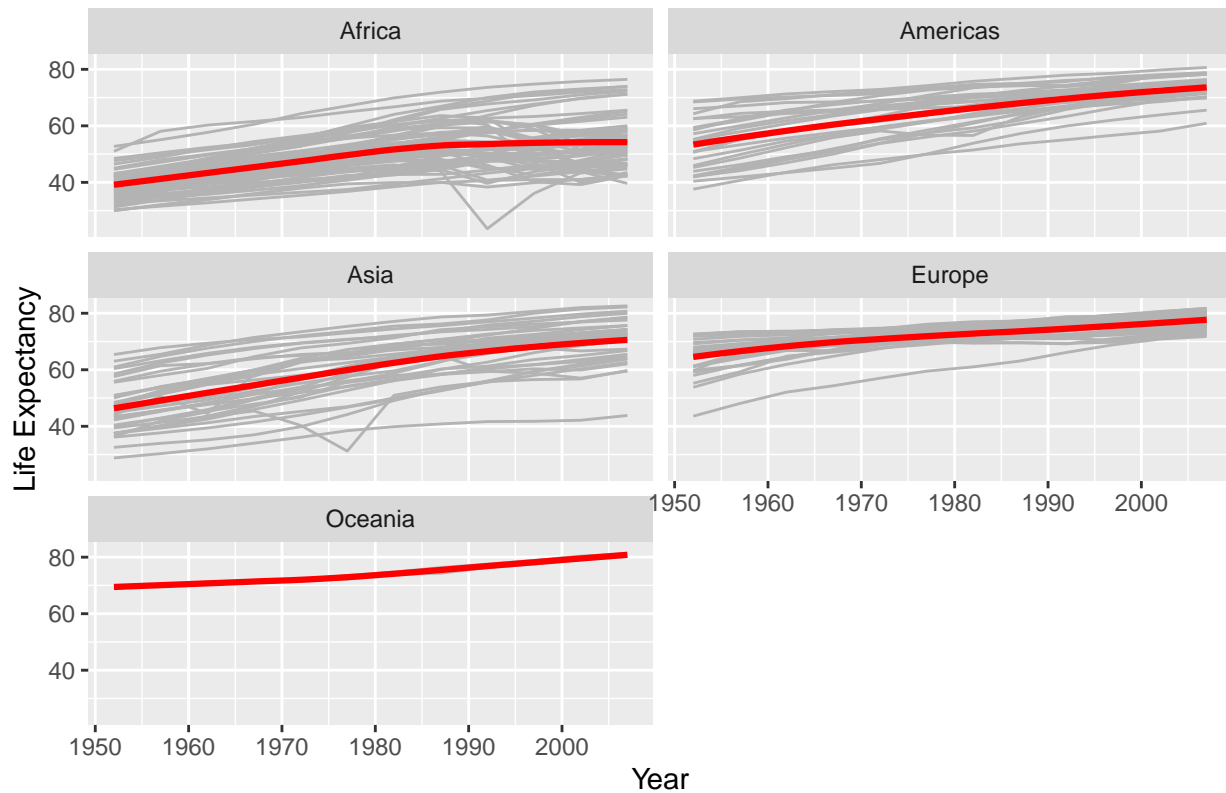
```
ggplot(data = gapminder,
       mapping = aes(
         x = year,
         y = lifeExp,
         group = country)) +
  geom_line() + facet_wrap(~ continent)
```



```
ggplot(data = gapminder,
       mapping = aes(
         x = year,
         y = lifeExp)) +
  geom_line(mapping = aes(group = country), color = "gray70") +
  geom_smooth(method = "loess", color="red", se = FALSE, size = 1.1) +
  facet_wrap(~ continent, ncol = 2) +
  labs(
    x = "Year",
    y = "Life Expectancy",
    title = " The trend of life expectancy of countries across 5 continents" )
```

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

## The trend of life expectancy of countries across 5 continents



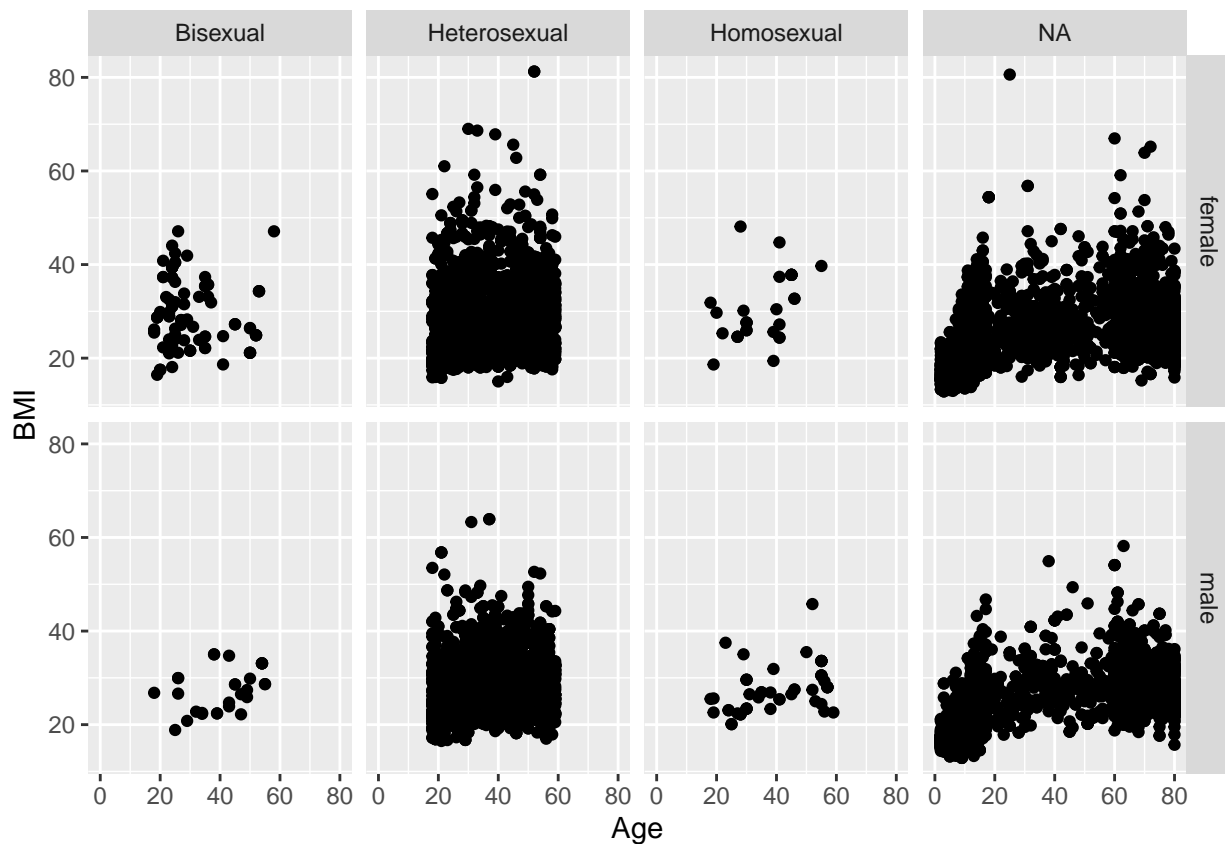
```
library("NHANES")
data(NHANES)
head(NHANES)
```

```
## # A tibble: 6 x 76
##   ID SurveyYr Gender   Age AgeDecade AgeMonths Race1 Race3 Education
##   <int> <fct>   <fct>   <int> <fct>      <int> <fct> <fct> <fct>
## 1 51624 2009_10 male     34 " 30-39"    409 White <NA> High Sch~
## 2 51624 2009_10 male     34 " 30-39"    409 White <NA> High Sch~
## 3 51624 2009_10 male     34 " 30-39"    409 White <NA> High Sch~
## 4 51625 2009_10 male      4 " 0-9"      49 Other <NA> <NA>
## 5 51630 2009_10 female   49 " 40-49"   596 White <NA> Some Col~
## 6 51638 2009_10 male      9 " 0-9"     115 White <NA> <NA>
## # ... with 67 more variables: MaritalStatus <fct>, HHIncome <fct>,
## #   HHIncomeMid <int>, Poverty <dbl>, HomeRooms <int>, HomeOwn <fct>,
## #   Work <fct>, Weight <dbl>, Length <dbl>, HeadCirc <dbl>, Height <dbl>,
## #   BMI <dbl>, BMICatUnder20yrs <fct>, BMI_WHO <fct>, Pulse <int>,
## #   BPSysAve <int>, BPDiaAve <int>, BPSys1 <int>, BPDial <int>, BPSys2 <int>,
## #   BPDia2 <int>, BPSys3 <int>, BPDia3 <int>, Testosterone <dbl>,
## #   DirectChol <dbl>, TotChol <dbl>, UrineVol1 <int>, UrineFlow1 <dbl>,
## #   UrineVol2 <int>, UrineFlow2 <dbl>, Diabetes <fct>, DiabetesAge <int>,
## #   HealthGen <fct>, DaysPhysHlthBad <int>, DaysMentHlthBad <int>,
## #   LittleInterest <fct>, Depressed <fct>, nPregnancies <int>, nBabies <int>,
## #   Age1stBaby <int>, SleepHrsNight <int>, SleepTrouble <fct>,
## #   PhysActive <fct>, PhysActiveDays <int>, TVHrsDay <fct>, CompHrsDay <fct>,
## #   TVHrsDayChild <int>, CompHrsDayChild <int>, Alcohol12PlusYr <fct>,
```

```
## # AlcoholDay <int>, AlcoholYear <int>, SmokeNow <fct>, Smoke100 <fct>,
## # Smoke100n <fct>, SmokeAge <int>, Marijuana <fct>, AgeFirstMarij <int>,
## # RegularMarij <fct>, AgeRegMarij <int>, HardDrugs <fct>, SexEver <fct>,
## # SexAge <int>, SexNumPartnLife <int>, SexNumPartYear <int>, SameSex <fct>,
## # SexOrientation <fct>, PregnantNow <fct>
```

```
ggplot(data = NHANES,
       mapping = aes(
         x = Age,
         y = BMI)) +
  geom_point() +
  facet_grid(Gender~ SexOrientation)
```

```
## Warning: Removed 366 rows containing missing values (geom_point).
```

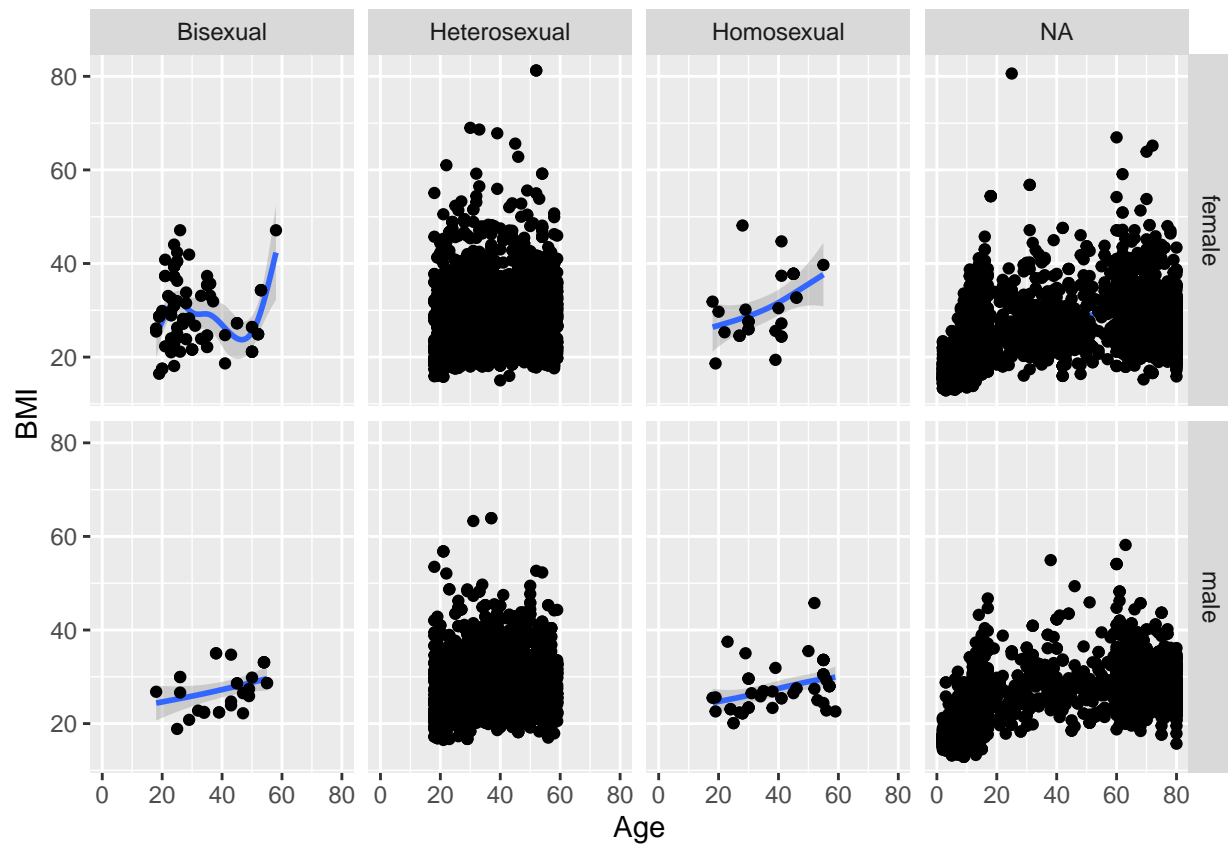


```
ggplot(data = NHANES,
       mapping = aes(
         x = Age,
         y = BMI)) +
  geom_smooth() +
  geom_point() +
  facet_grid(Gender~ SexOrientation)
```

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

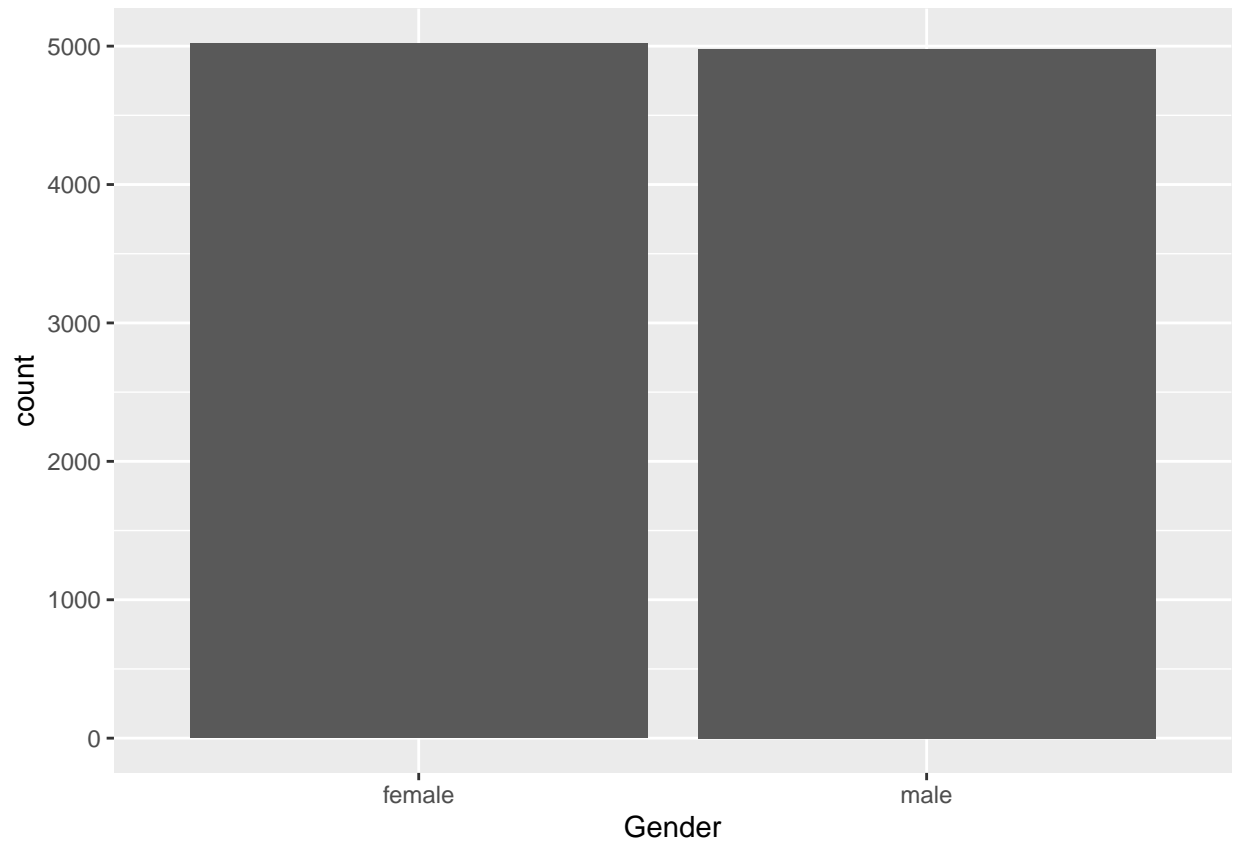
```
## Warning: Removed 366 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 366 rows containing missing values (geom_point).
```

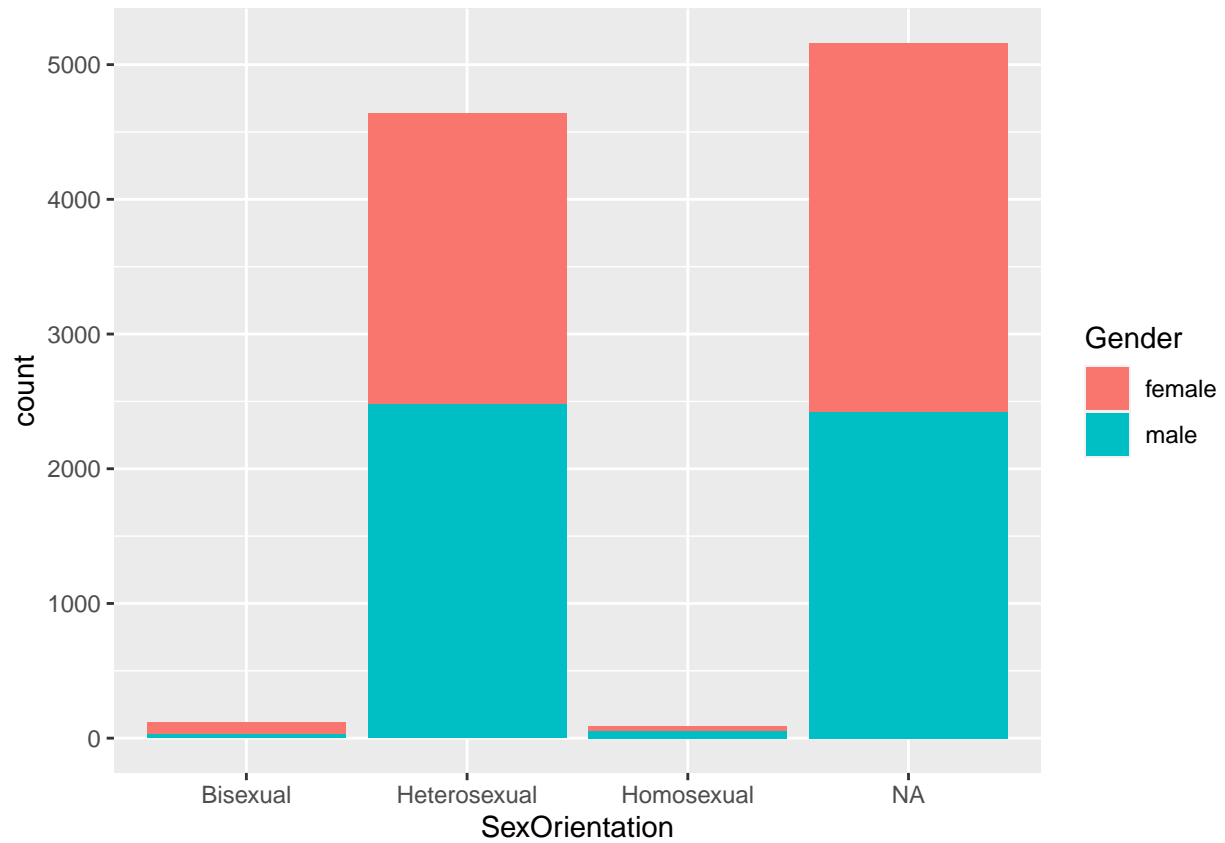


## 4 Barplots

```
ggplot(data = NHANES,
       mapping = aes(x = Gender)) +
  geom_bar()
```

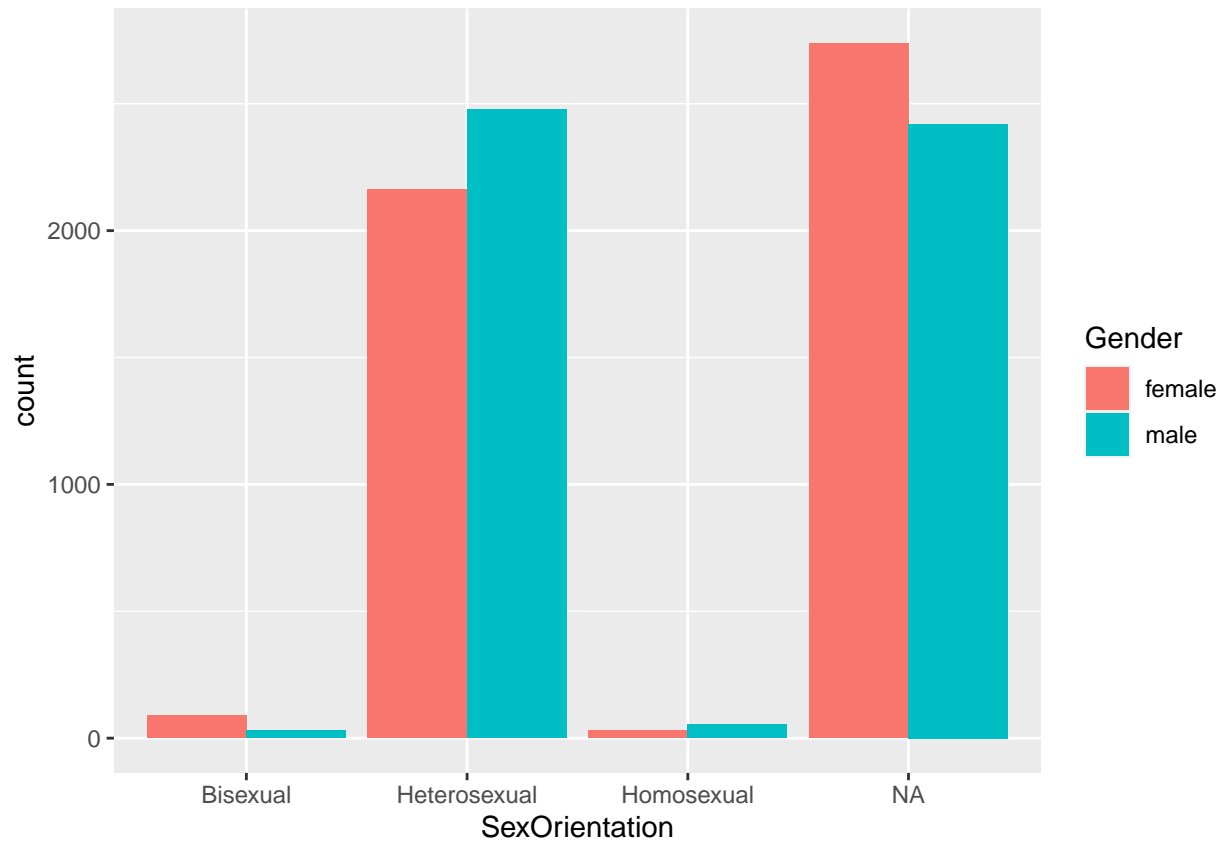


```
ggplot(data = NHANES,  
       mapping = aes(x = SexOrientation, fill = Gender)) +  
  geom_bar()
```

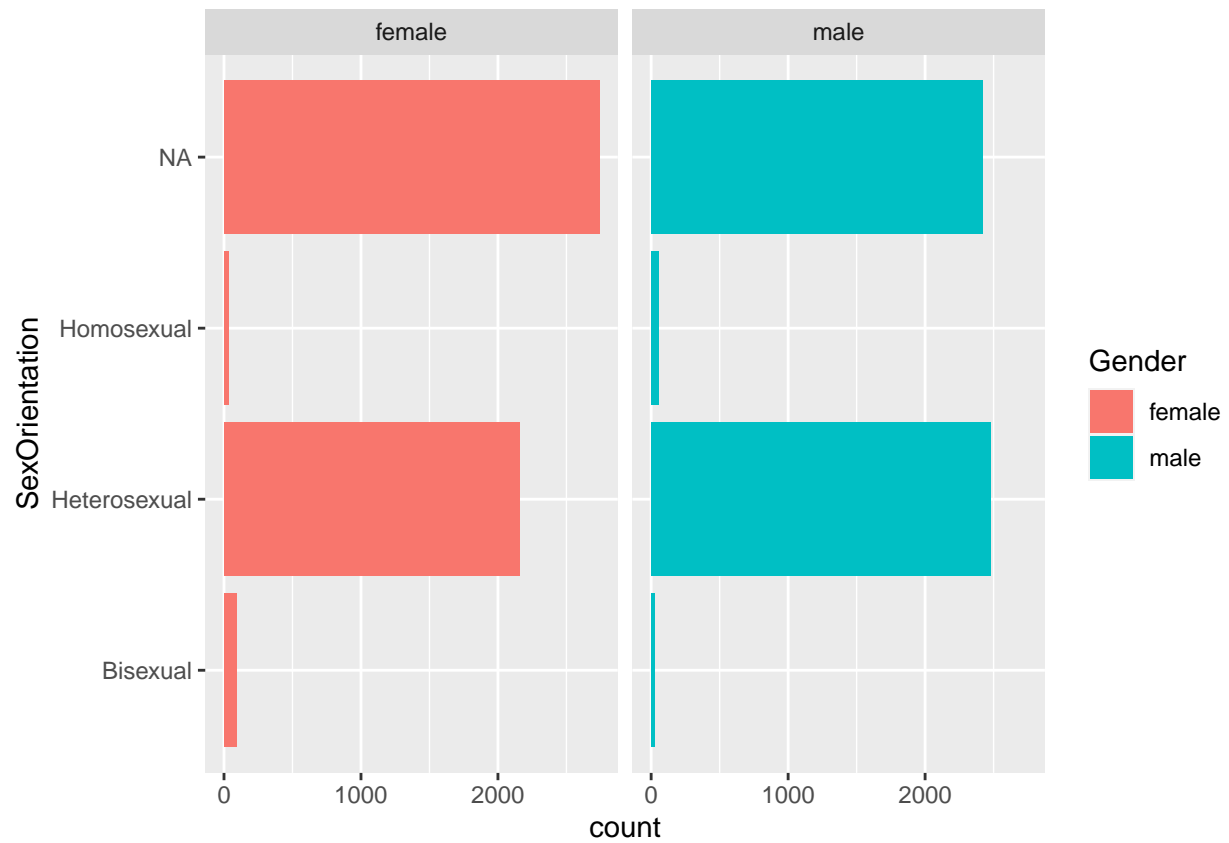


```
ggplot(data = NHANES,  
  mapping = aes(x = SexOrientation, fill = Gender)) +  
  geom_bar(position = "dodge") # stack
```

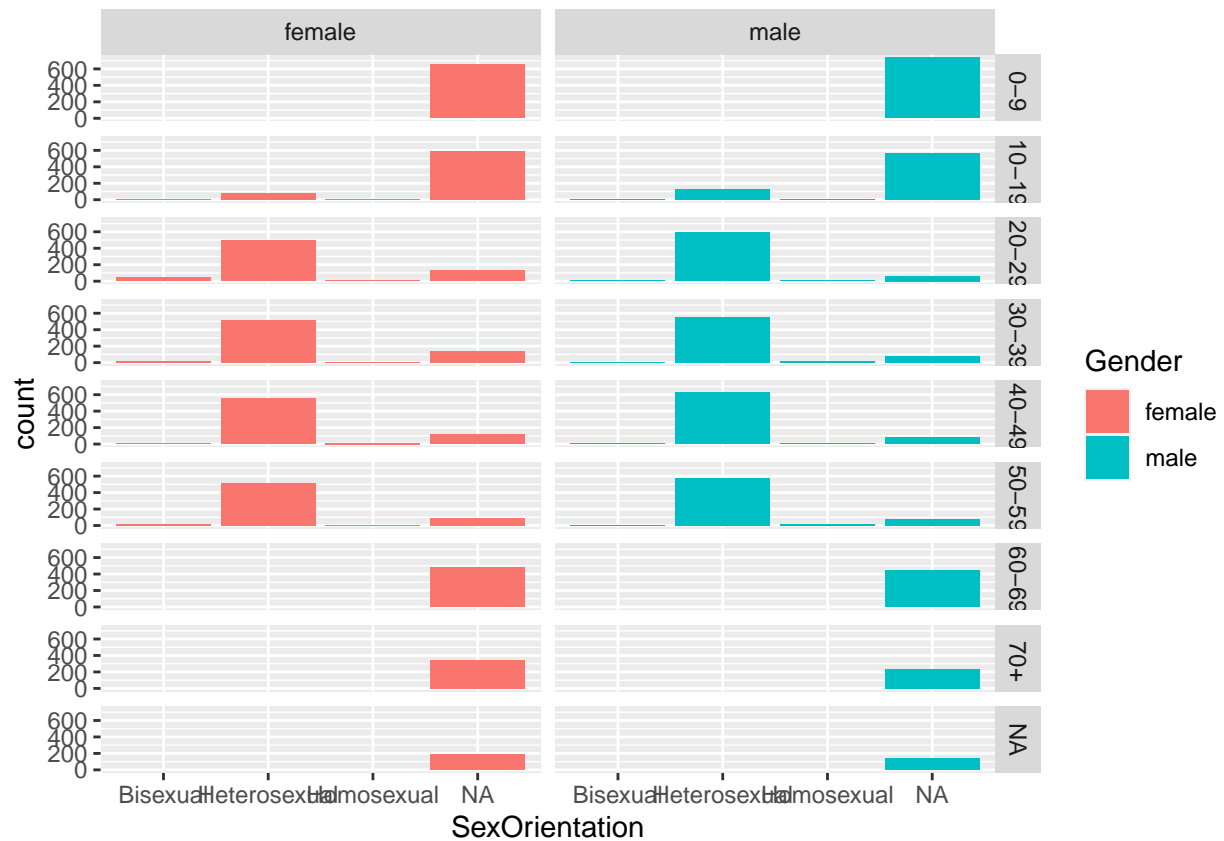




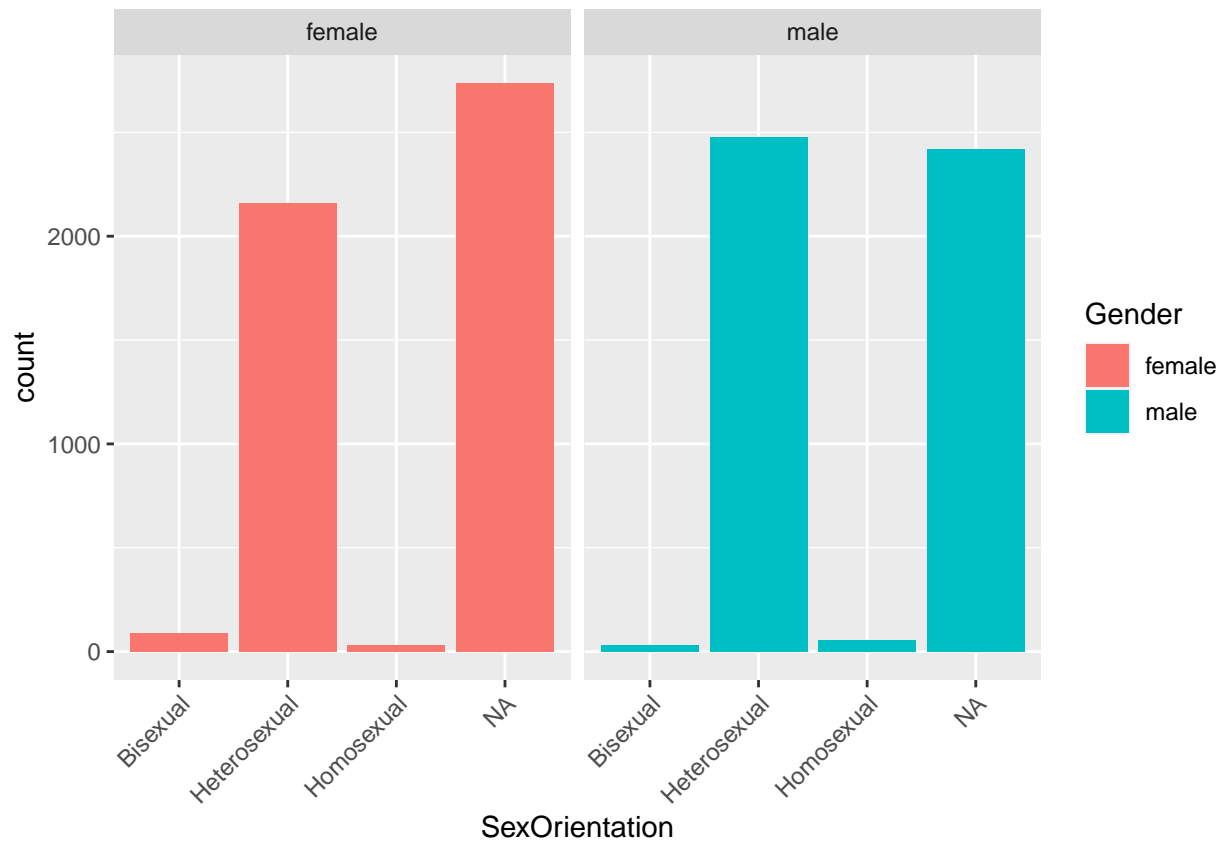
```
ggplot(data = NHANES,  
       mapping = aes(x = SexOrientation, fill = Gender)) +  
  geom_bar(position = "dodge") +  
  coord_flip() +  
  facet_grid(~ Gender)
```



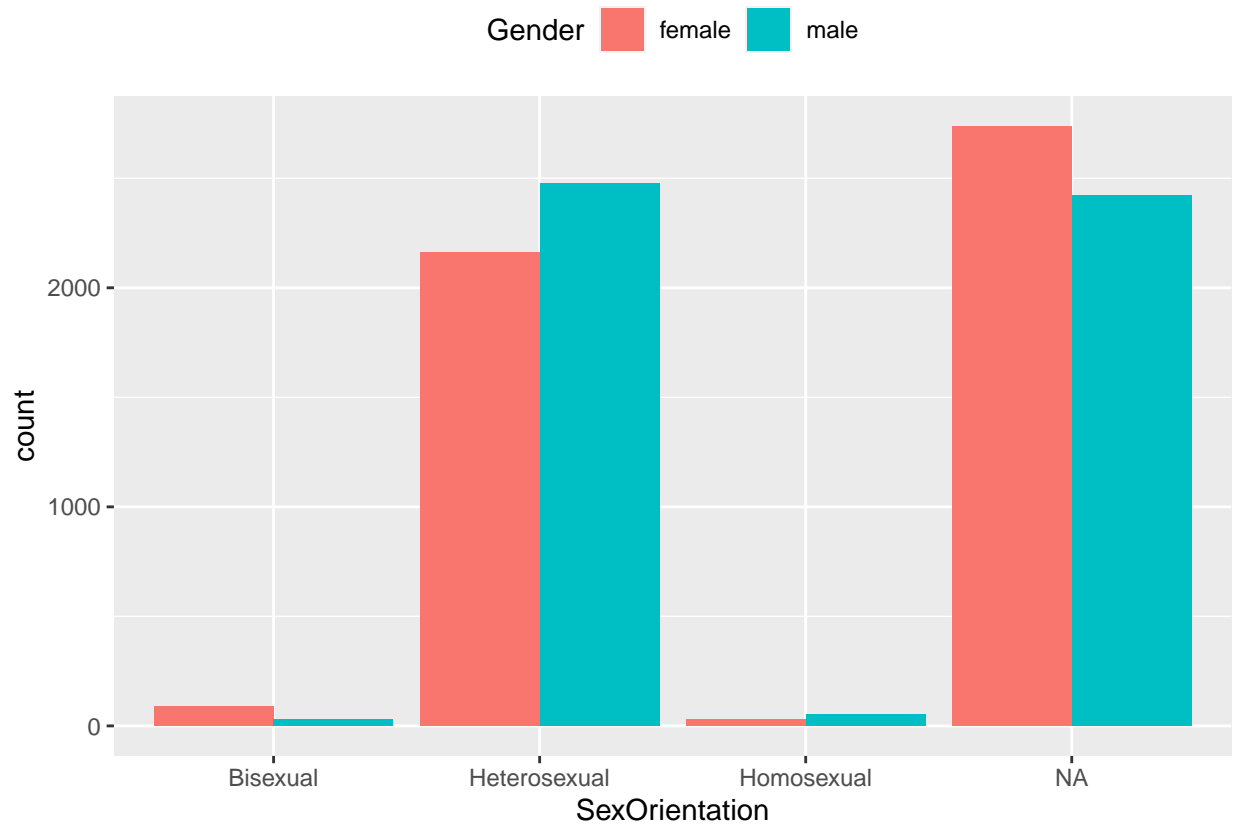
```
ggplot(data = NHANES,
       mapping = aes(x = SexOrientation, fill = Gender)) +
  geom_bar(position = "dodge") +
  facet_grid(AgeDecade ~ Gender)
```



```
ggplot(data = NHANES,
       mapping = aes(x = SexOrientation, fill = Gender)) +
  geom_bar(position = "dodge") +
  facet_grid(~ Gender) +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1) )
```



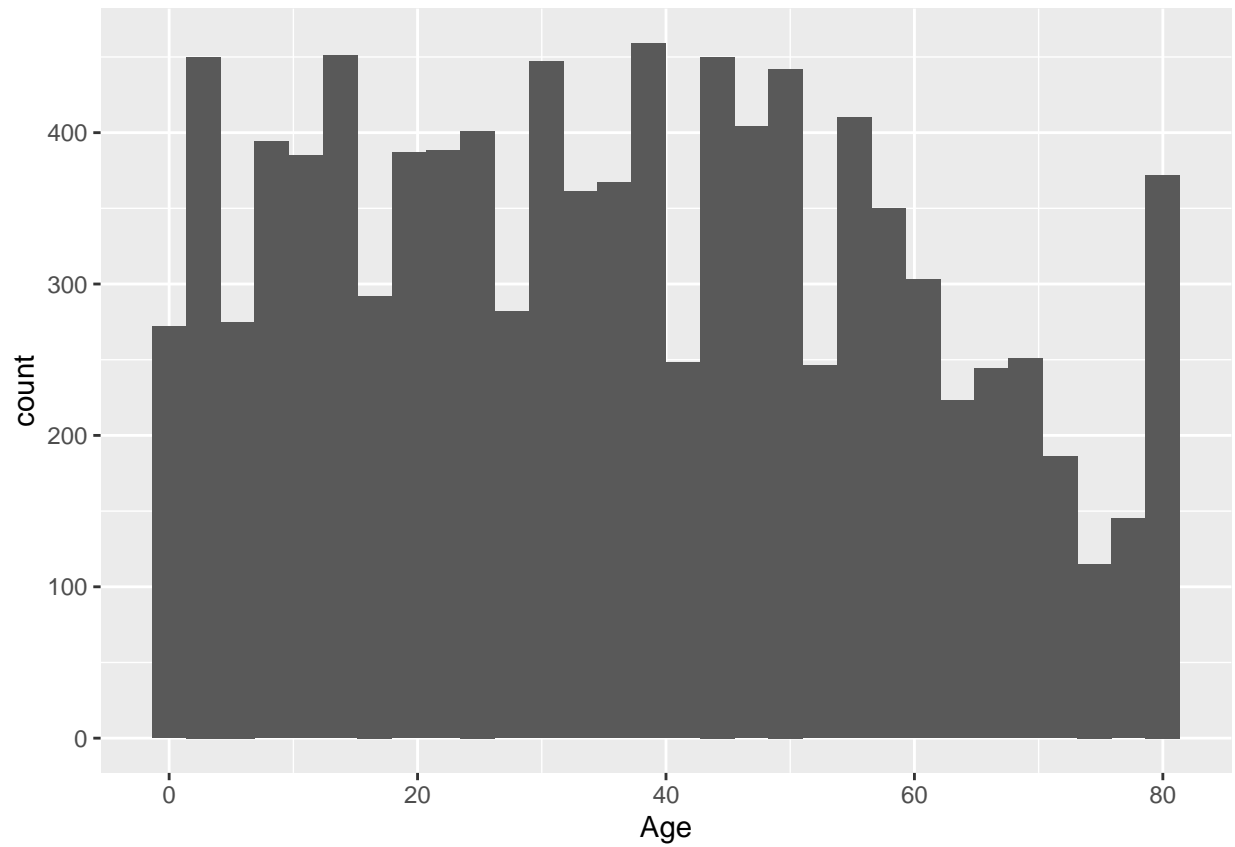
```
# change the legends  
ggplot(data = NHANES,  
       mapping = aes(x = SexOrientation, fill = Gender)) +  
  geom_bar(position = "dodge") +  
  theme(legend.position = "top") # top left right
```



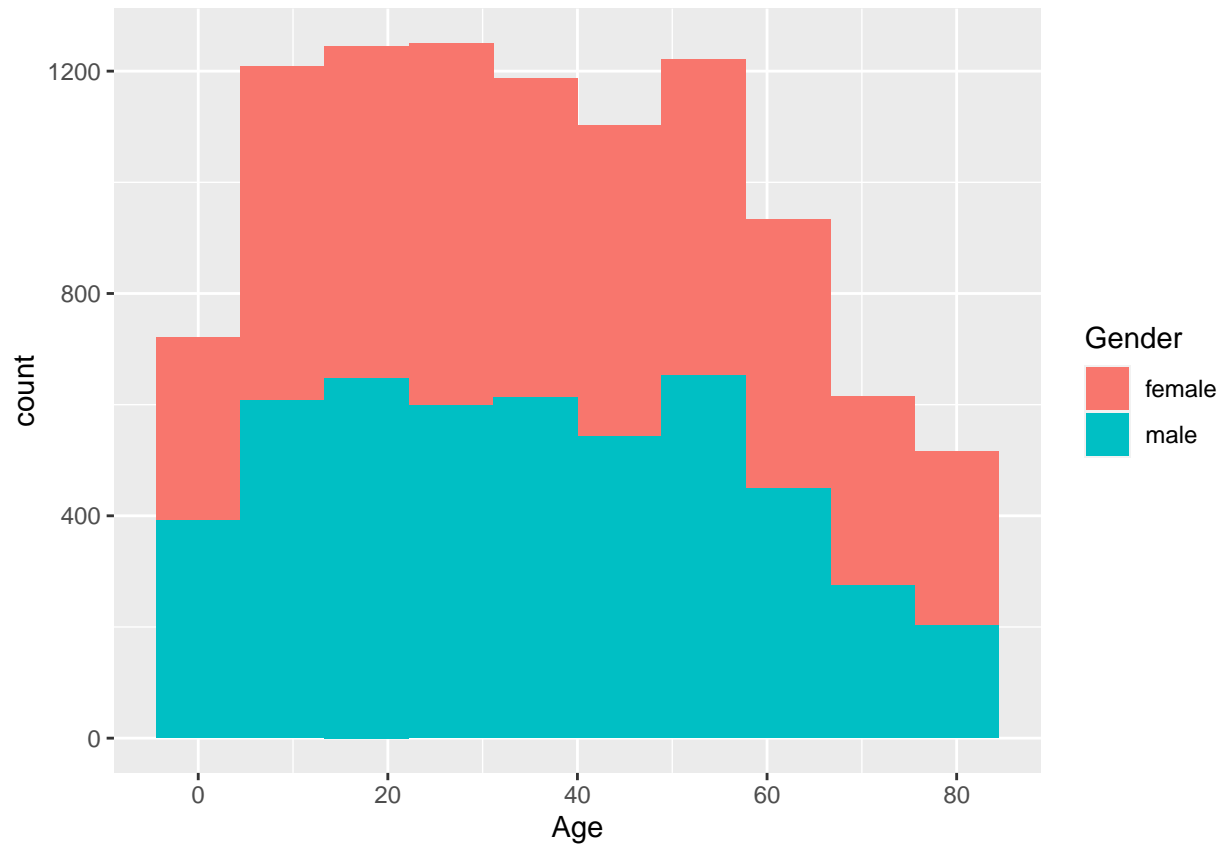
## 5 Histograms

```
ggplot(data = NHANES,  
       mapping = aes(x = Age)) +  
  geom_histogram() # bins = 20
```

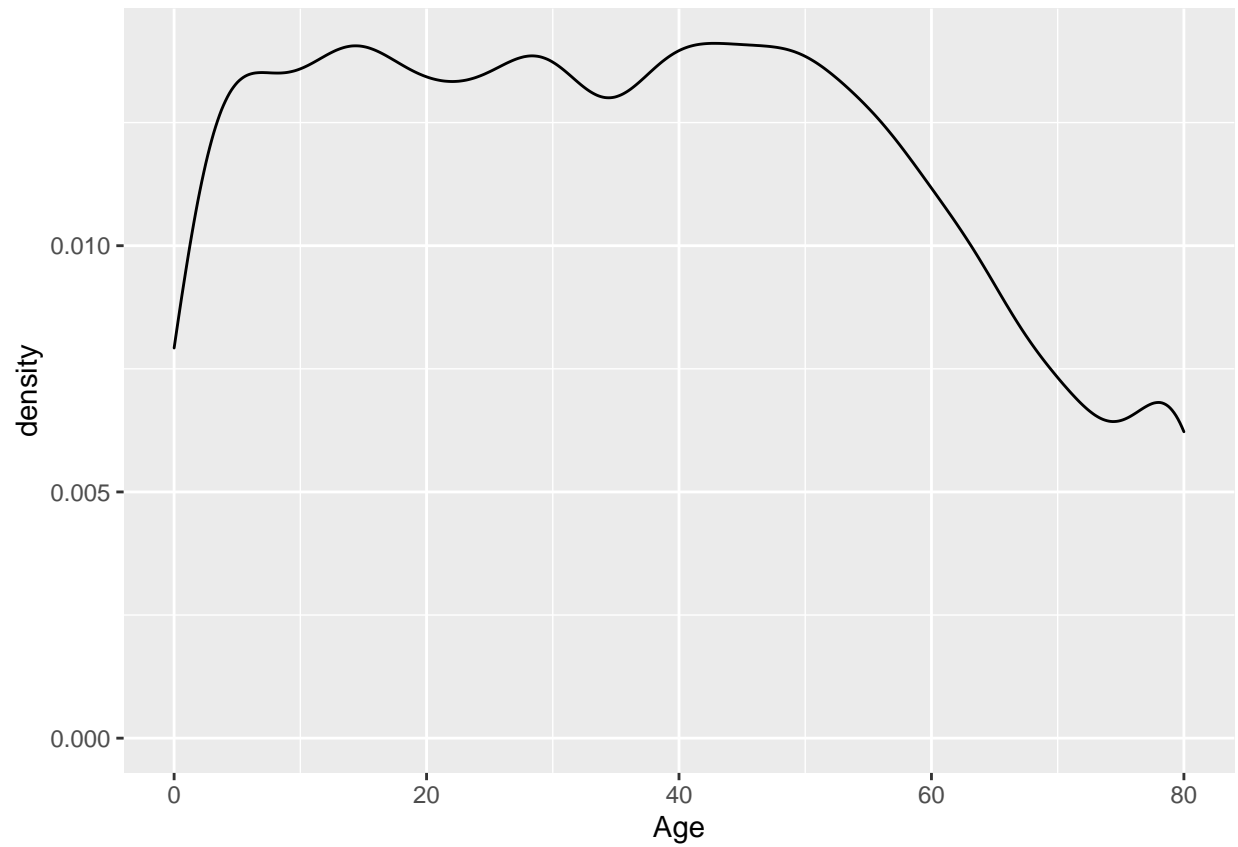
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(data = NHANES,  
       mapping = aes(x = Age, fill = Gender)) +  
  geom_histogram(bins = 10)
```

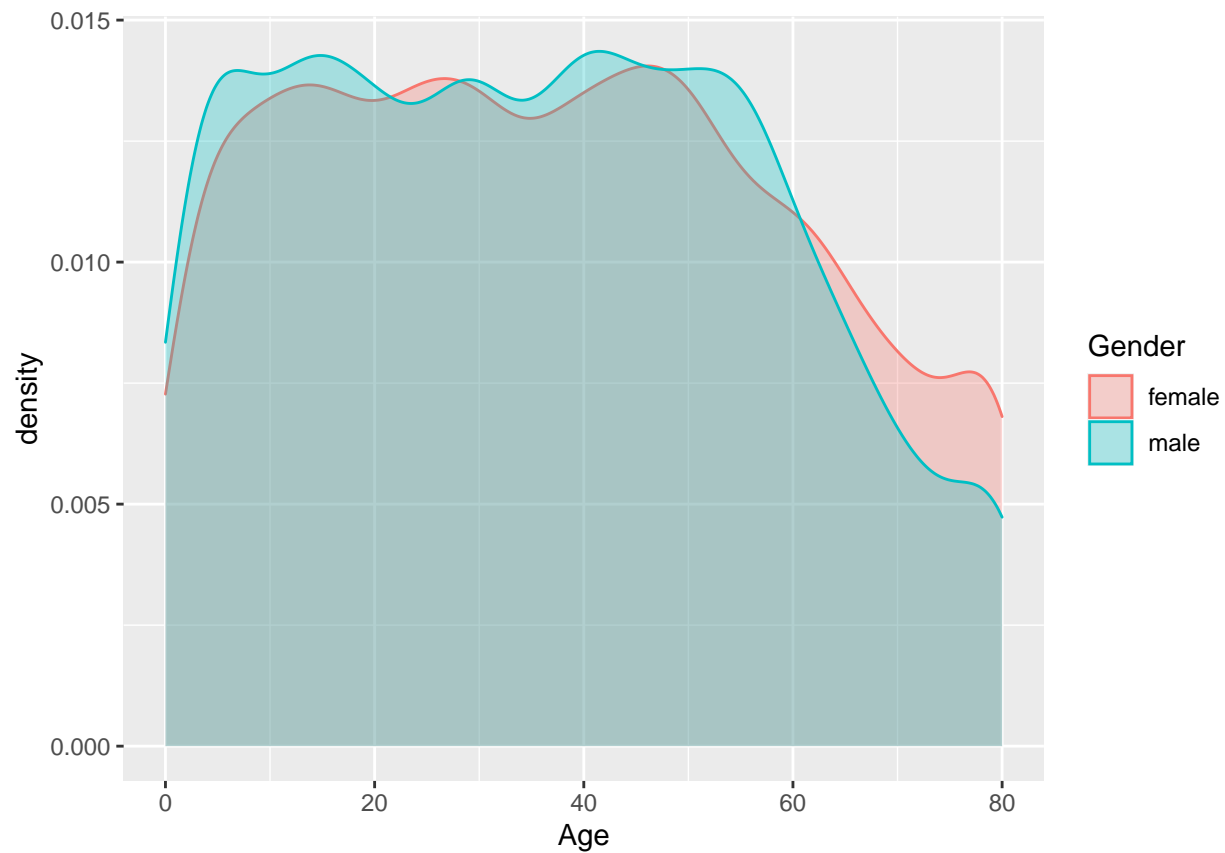


```
ggplot(data = NHANES,  
       mapping = aes(x = Age)) +  
  geom_density()
```

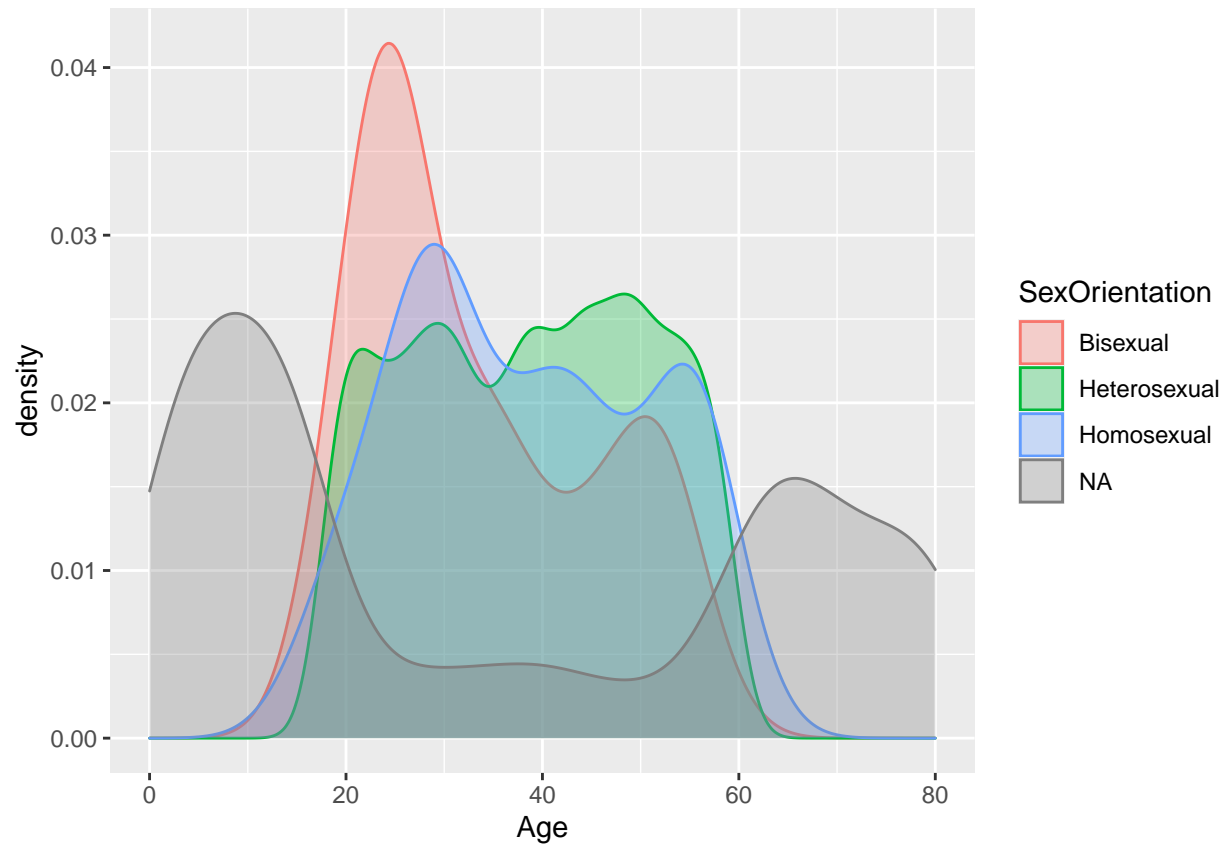


```
ggplot(data = NHANES,  
       mapping = aes(x = Age, fill = Gender, color = Gender)) +  
  geom_density(alpha = 0.3)
```

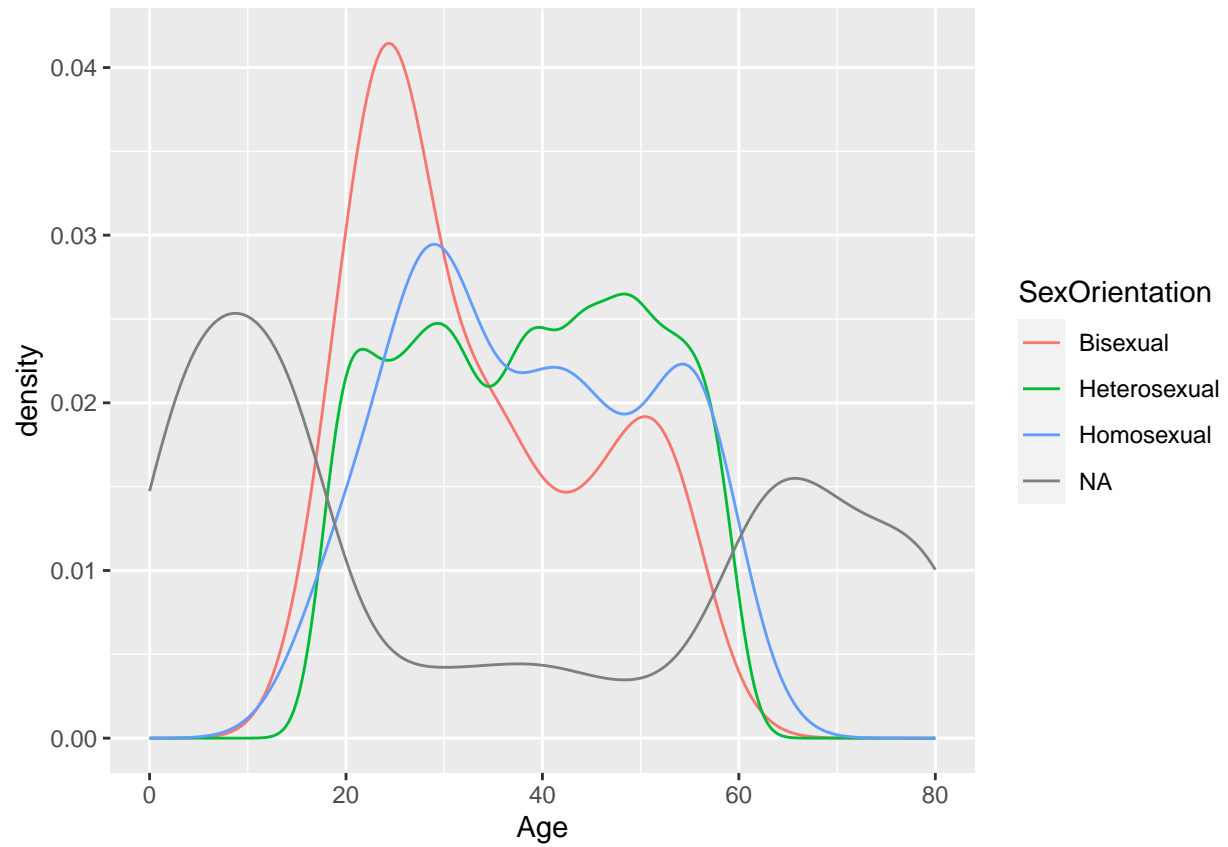




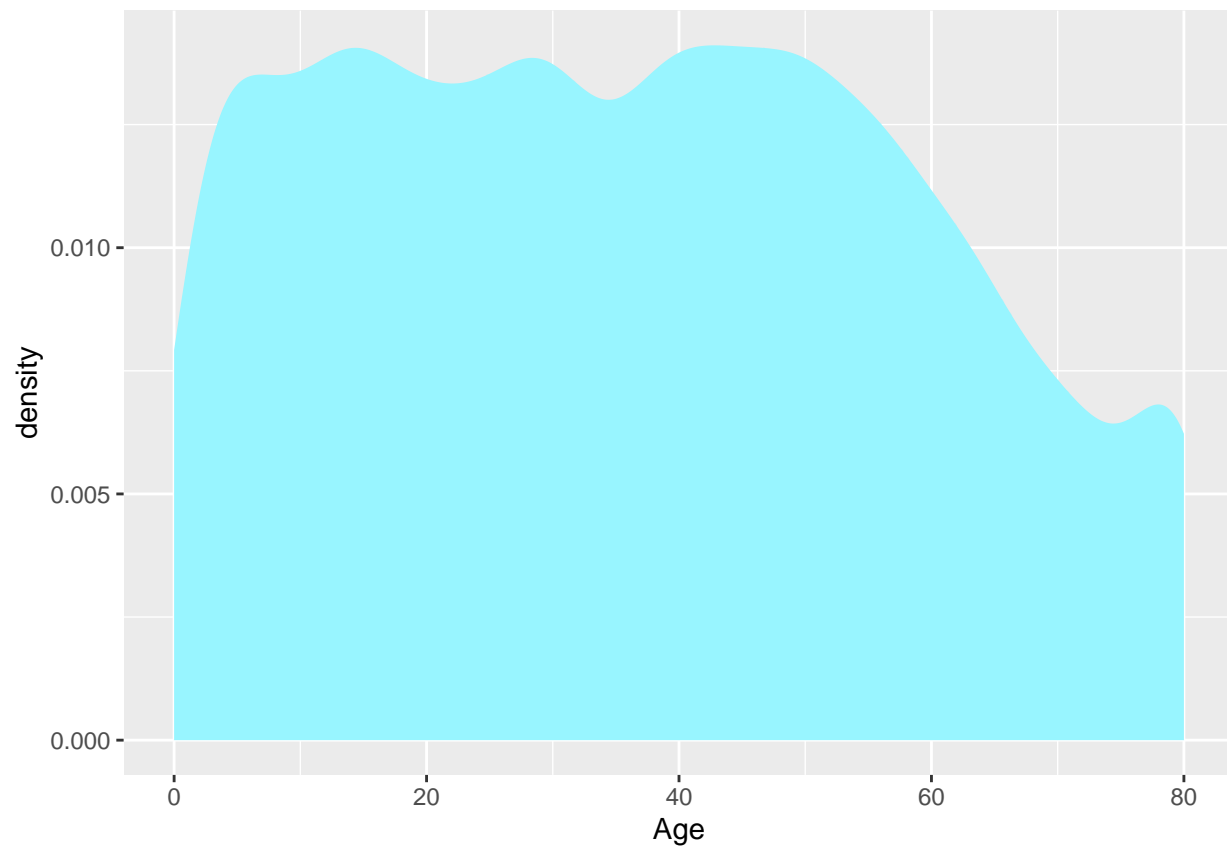
```
ggplot(data = NHANES,  
       mapping = aes(x = Age, fill = SexOrientation, color = SexOrientation)) +  
  geom_density(alpha = 0.3)
```



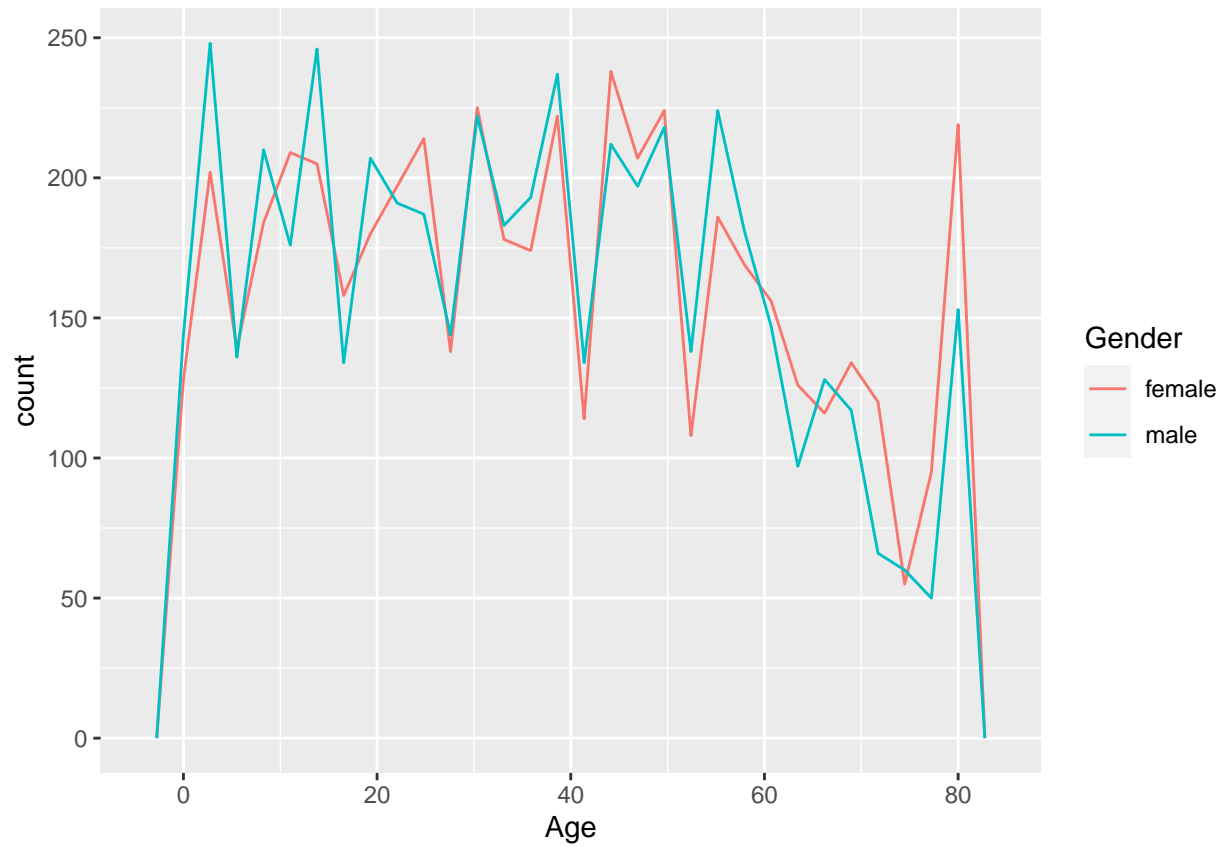
```
# add the density curves  
ggplot(data = NHANES,  
       mapping = aes(x = Age, color = SexOrientation)) +  
  geom_line(stat = "density")
```



```
ggplot(data = NHANES,  
  mapping = aes(x = Age)) +  
  geom_area(stat = "density", fill = "cadetblue1")
```

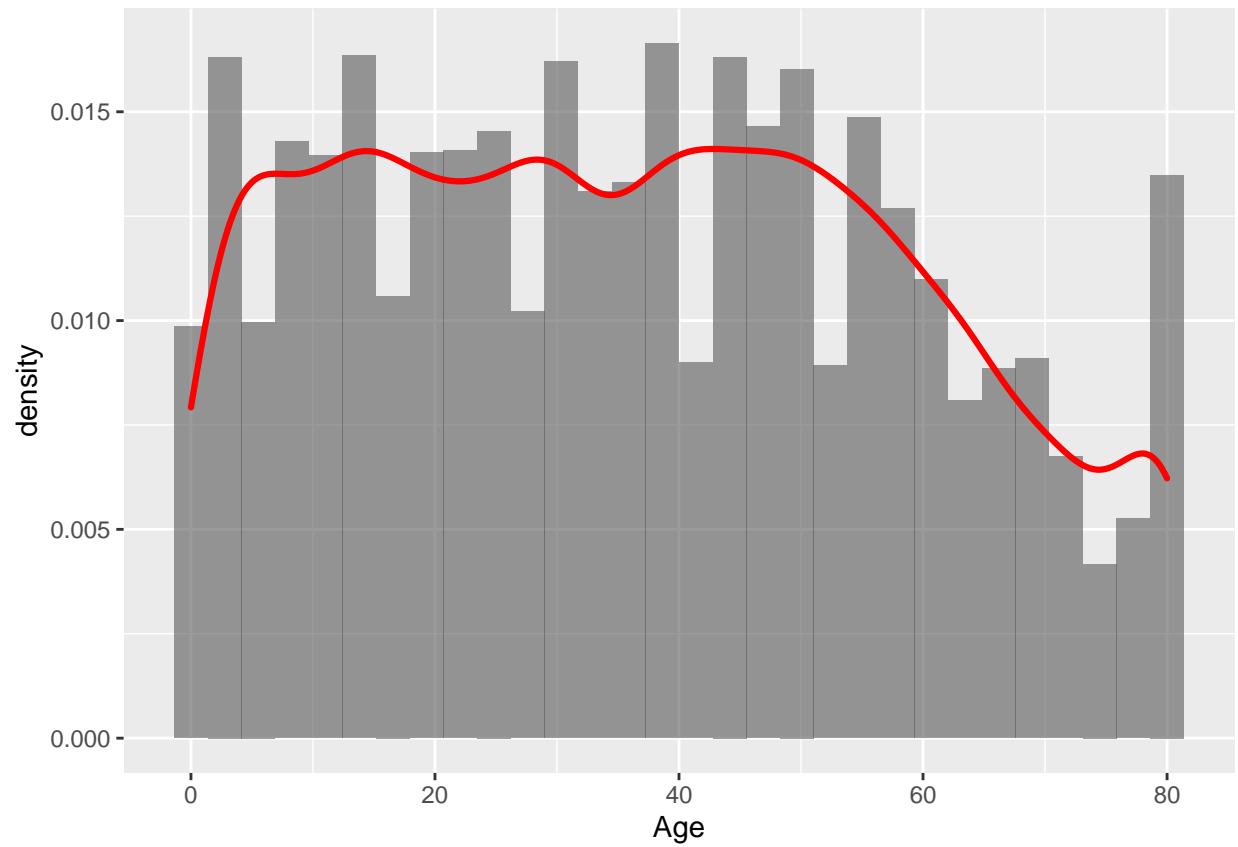


```
ggplot(data = NHANES,  
       mapping = aes(x = Age)) +  
geom_freqpoly(mapping = aes(color = Gender))  
  
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



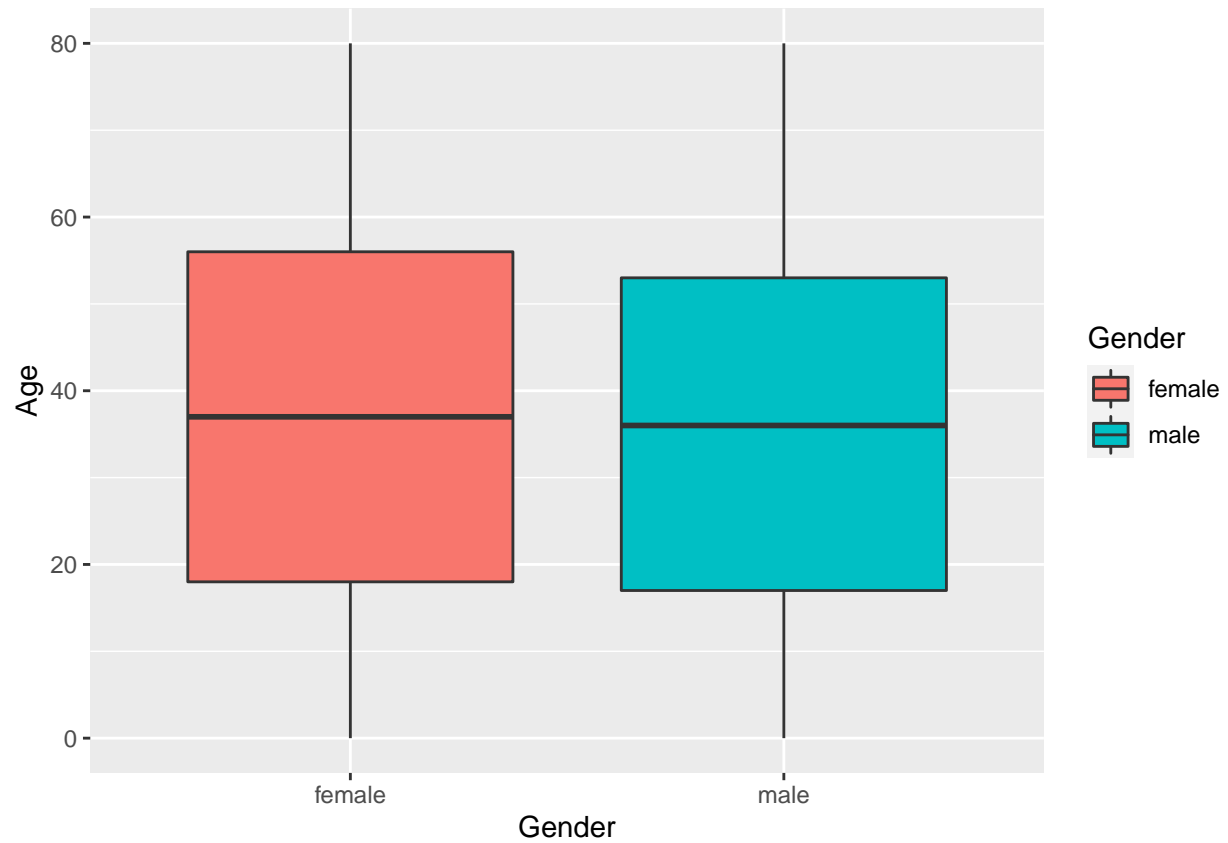
```
ggplot(data = NHANES, mapping = aes(x = Age)) +  
  geom_histogram(mapping = aes(y = ..density..), alpha = 0.6) +  
  geom_density(size = 1.1, color = "red")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

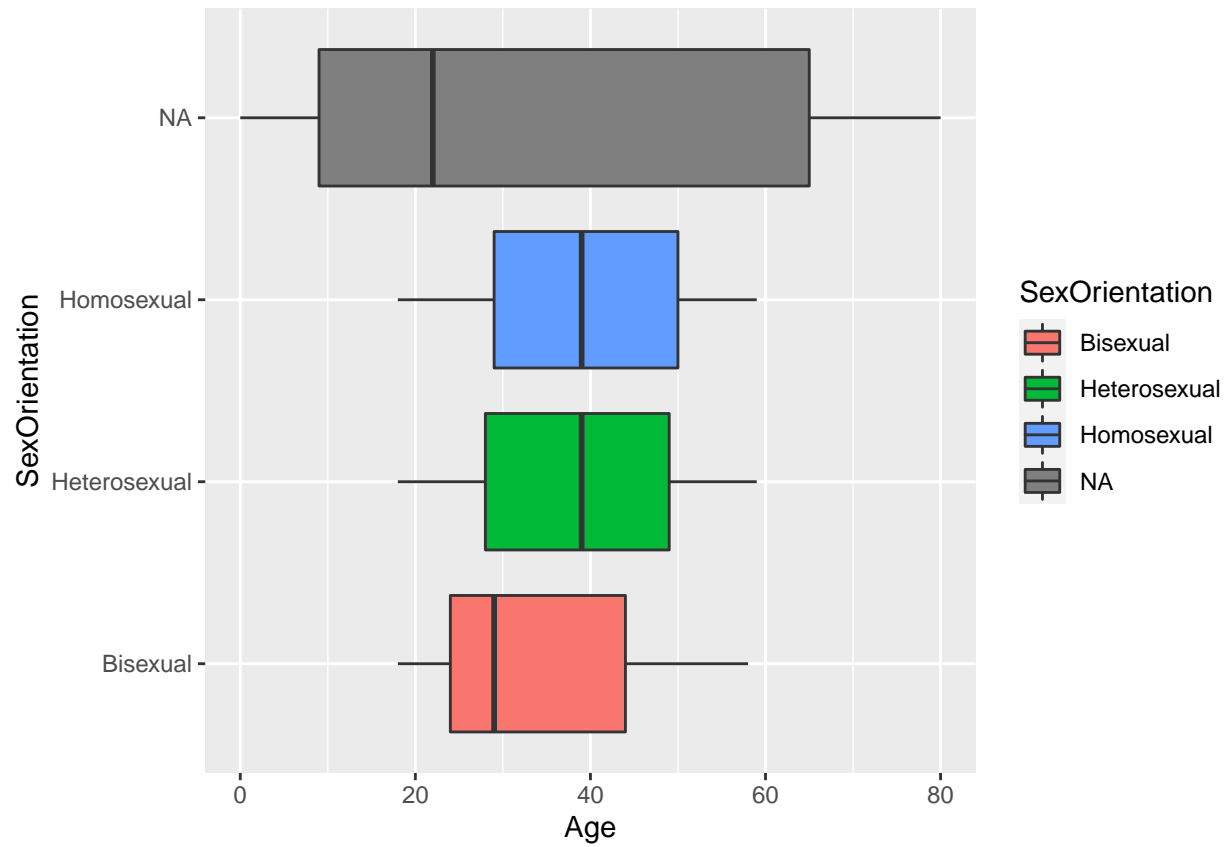


## 6 Boxplots

```
ggplot(data = NHANES,  
       mapping = aes(x = Gender, y = Age, fill = Gender)) +  
  geom_boxplot()
```



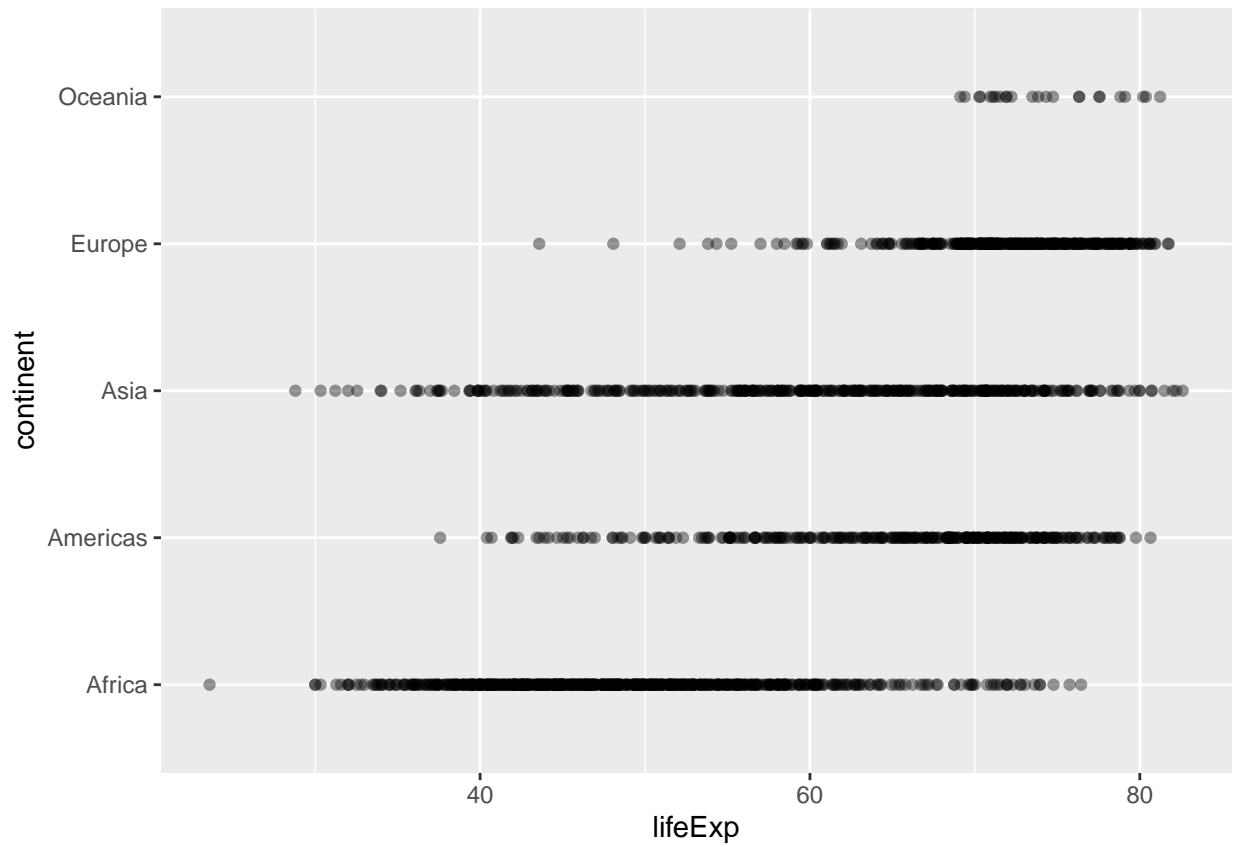
```
ggplot(data = NHANES,  
       mapping = aes(x = SexOrientation, y = Age, fill = SexOrientation)) +  
  geom_boxplot() +  
  coord_flip()
```



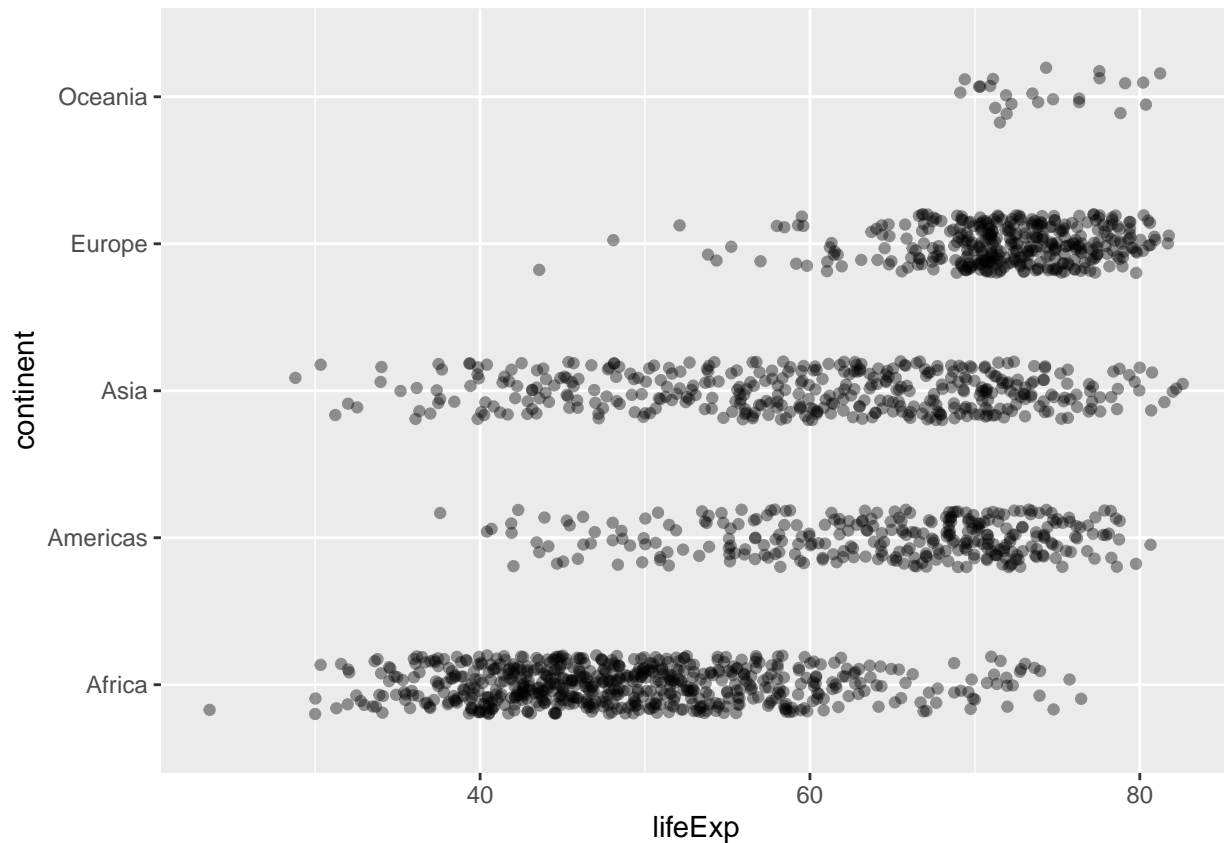
## 7 Dotted Plots

```
ggplot(data = gapminder,
       mapping = aes(
         y = lifeExp,
         x = continent)) +
  geom_point(alpha = 0.4) +
  coord_flip()
```





```
ggplot(data = gapminder,  
  mapping = aes(  
    y = lifeExp,  
    x = continent)) +  
  coord_flip() +  
  geom_jitter(alpha = 0.4, width = 0.2, height = 0)
```



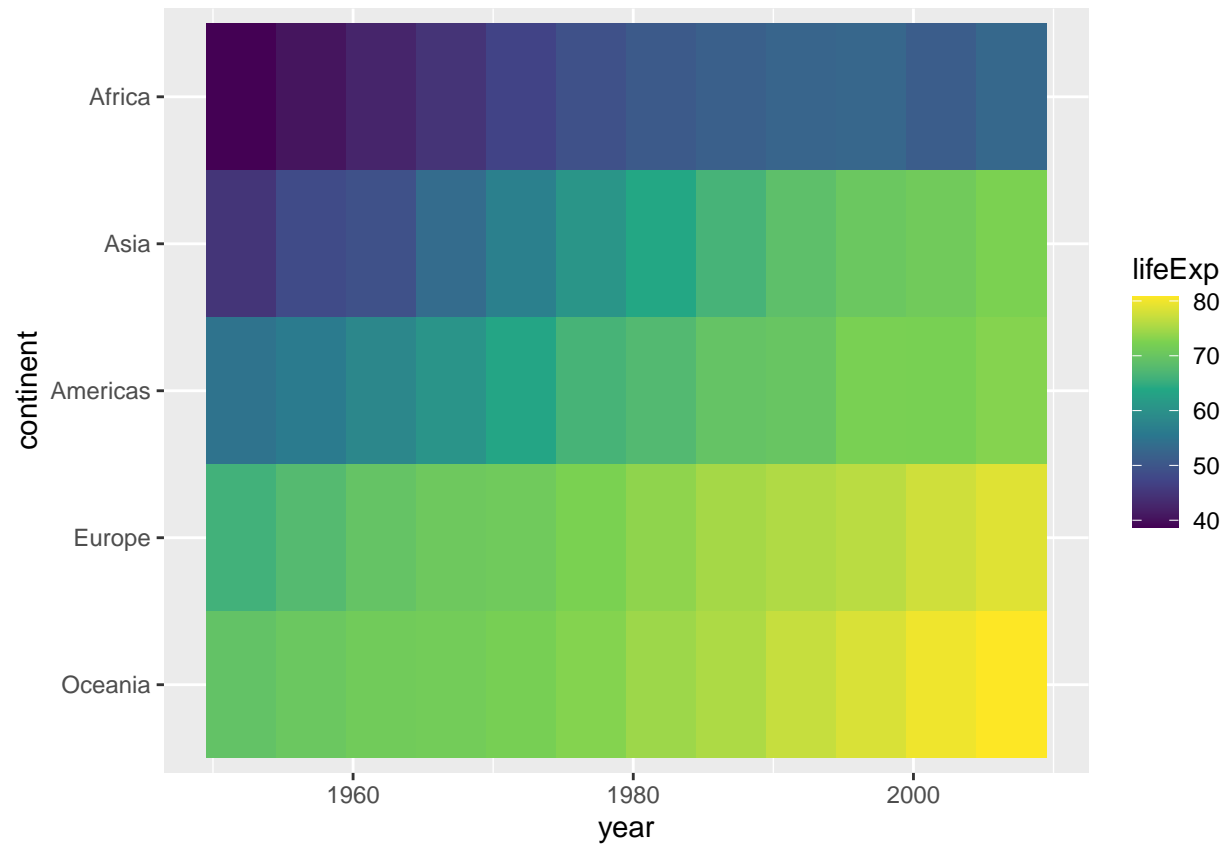
## 8 Heatmap

```
gapminder %>%
  select(continent, year, lifeExp) %>%
  group_by(continent, year) %>%
  summarise(lifeExp = median(lifeExp, na.rm=TRUE)) %>%
  ungroup() -> d_gap2a

## `summarise()` regrouping output by 'continent' (override with `.groups` argument)

d_gap2a %>%
  filter(year == 2007) %>%
  arrange(desc(lifeExp)) -> d_gap2b
d_gap2a %>%
  mutate(continent = factor(continent, levels = d_gap2b$continent)) -> d_gap2

ggplot(data = d_gap2, mapping = aes(
  x = year, y = continent, fill = lifeExp)) +
  geom_tile() +
  scale_fill_viridis_c()
```

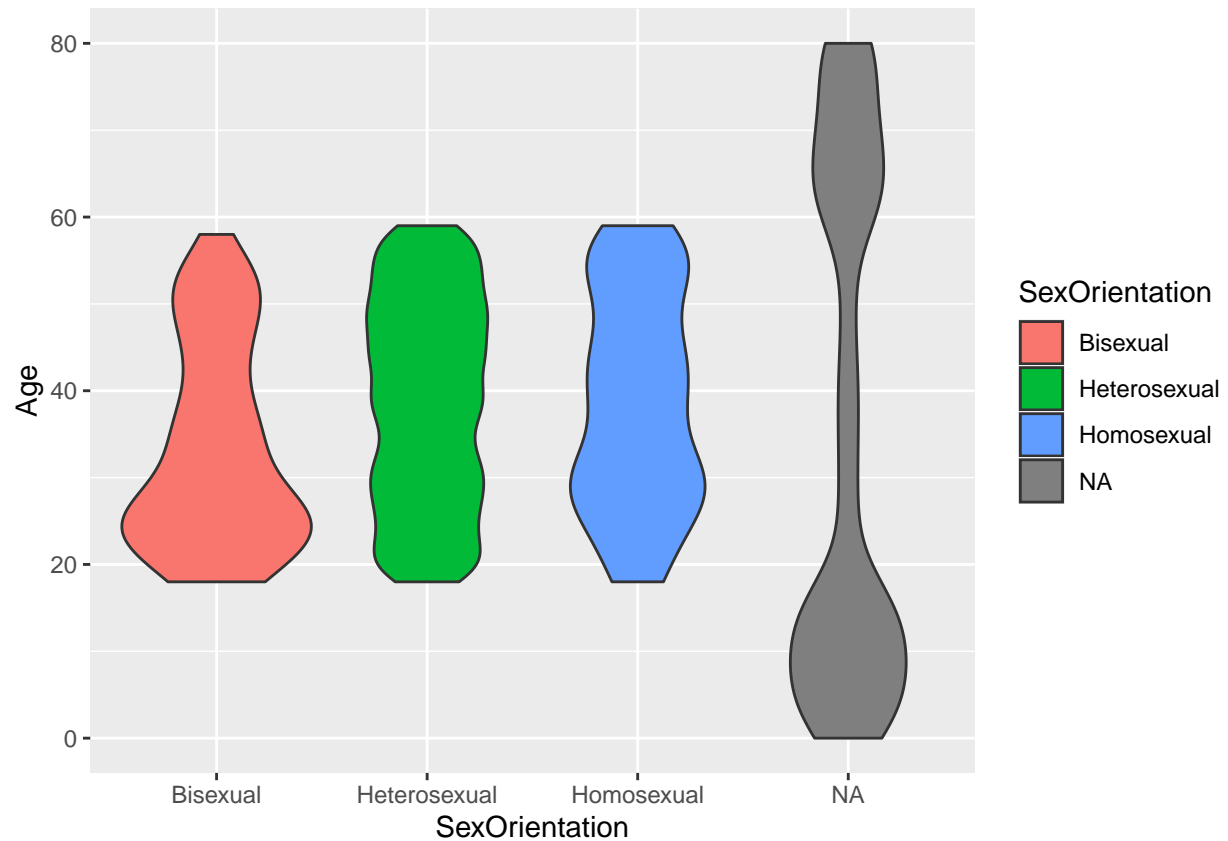


## 9 Violin Plots

```
ggplot(data = NHANES,  
       mapping = aes(x = Gender, y = Age, fill = Gender)) +  
  geom_violin()
```



```
ggplot(data = NHANES,  
       mapping = aes(x = SexOrientation, y = Age, fill = SexOrientation)) +  
  geom_violin()
```



## 10 References

Claus O. Wilke(2019). Fundamentals of Data Visualization. O'Reilly Media. <https://serialmentor.com/dataviz/>

Healy, Kieran (2018). Data Visualization: A Practical Introduction. Princeton University Press. <https://socviz.co/index.html>

Winston Chang(2018). R Graphics Cookbook. O'Reilly Media. <https://r-graphics.org/>