Introduction to Data Journalism

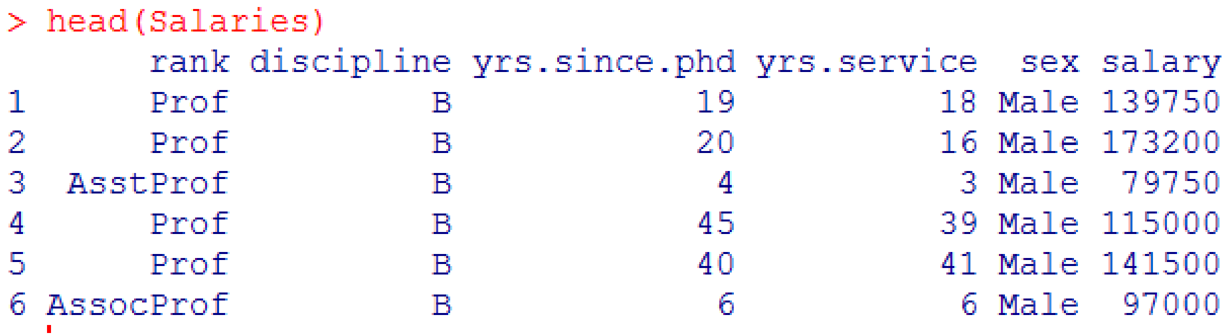
ggplot2

Datasets

data(Salaries, package="car")

Salaries from the **car** package
(2008-2009 9 month academic salaries n=397)

1. rank (AssocProf, AsstProf, Prof)
2. salary in dollars
3. discipline (A=theoretical, B=applied)
4. sex (Female, Male)
5. yrs.since.phd.
6. yrs.service



```
> head(Salaries)
       rank discipline yrs.since.phd yrs.service  sex salary
1      Prof          B            19          18 Male 139750
2      Prof          B            20          16 Male 173200
3  AsstProf          B             4           3 Male  79750
4      Prof          B            45          39 Male 115000
5      Prof          B            40          41 Male 141500
6 AssocProf          B             6           6 Male  97000
```

Grammar of Graphics

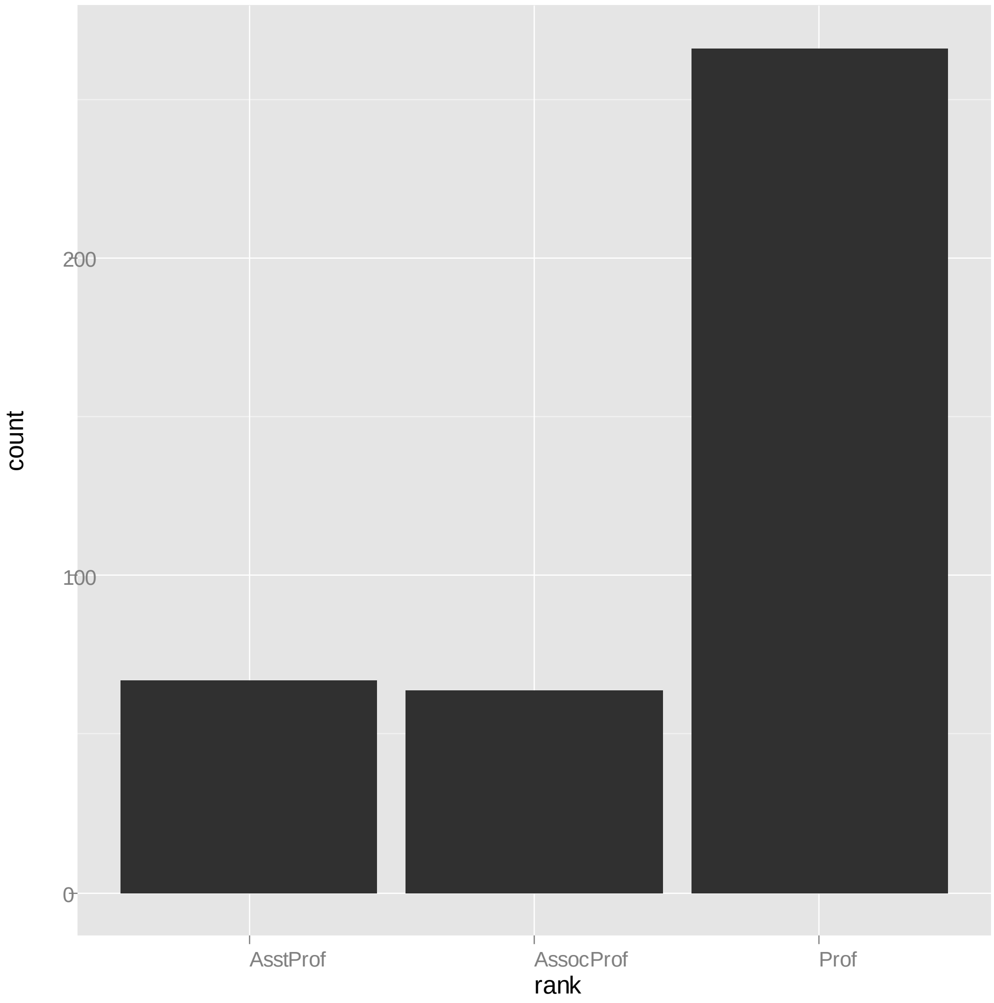
- data: an R data frame
- coordinate system: 2-D space data projected onto (e.g. cartesiona coordinates, polar coordinates, map projections)
- geoms: type of geometric objects that represent data (e.g. points, lines, bars)
- aesthetics: visual characteristics that represent data (e.g. position, size, color, shape, transparency, fill)
- scales: for each aesthetic, how visual characteristic is converted to display values
- stats: statistical transformations that summarize data (e.g., counts, means, trend lines)
- facets: how data is split into subsets and displayed as small multiples

Thanks to Dawn Koffman ggplot2 presentation

Simple bar plot

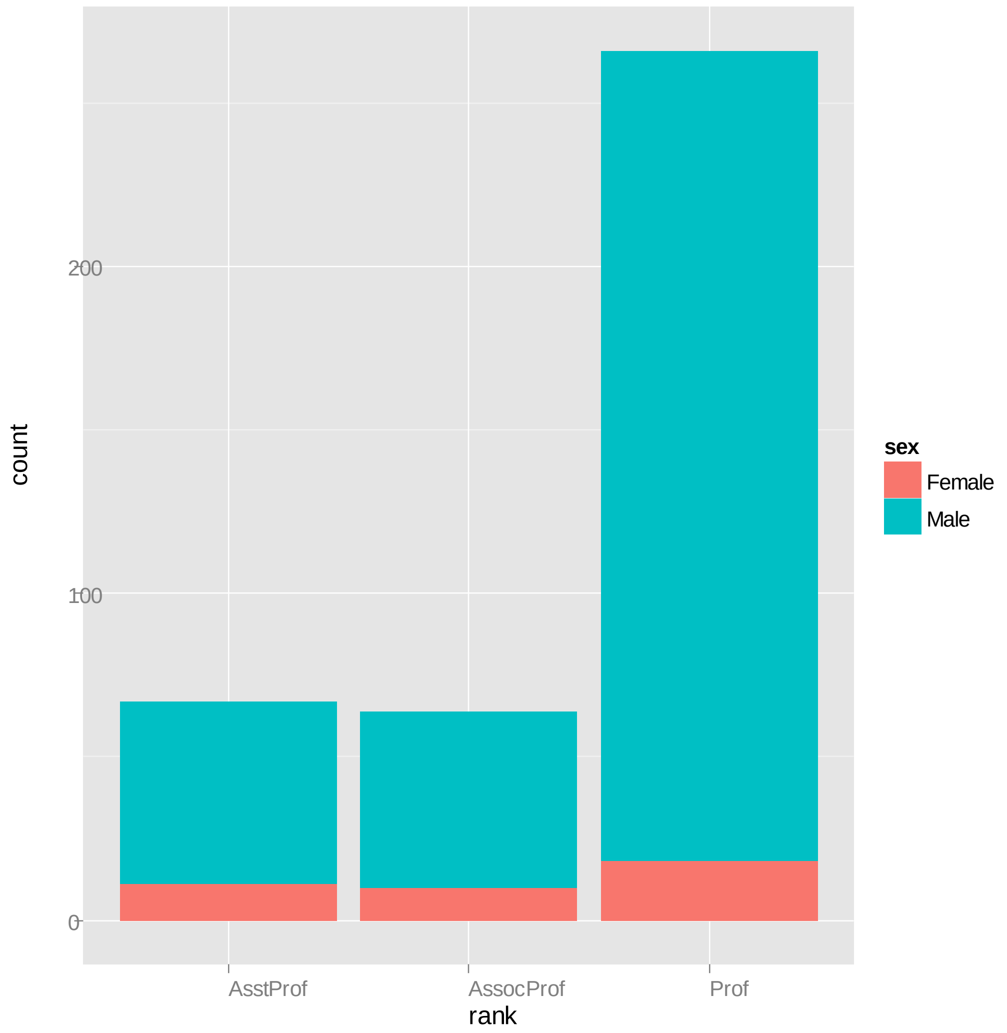
```
ggplot(data=Salaries,
aes(x=rank)) +
geom_bar()
```

common geom_bar options:
width
fill
color (border)
position_dodge()



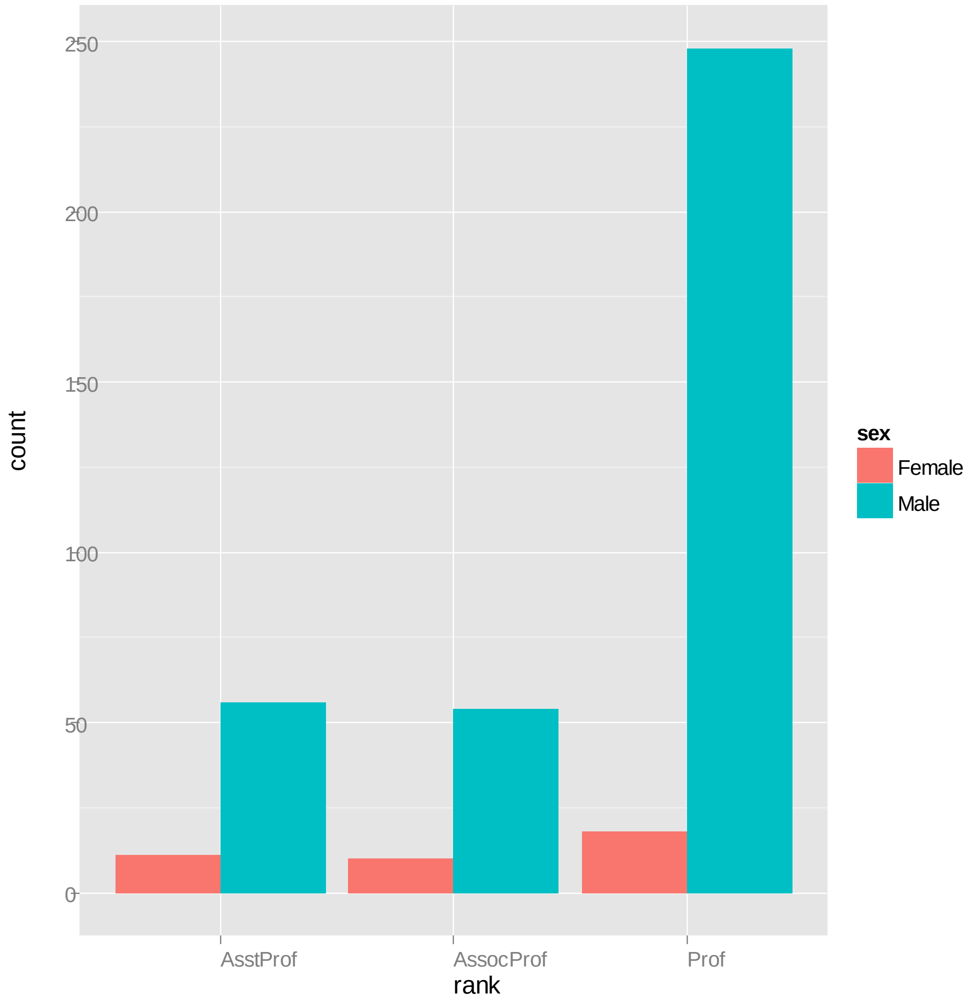
Stacked bar plot

```
ggplot(data=Salaries,
aes(x=rank, fill=sex)) +
geom_bar()
```



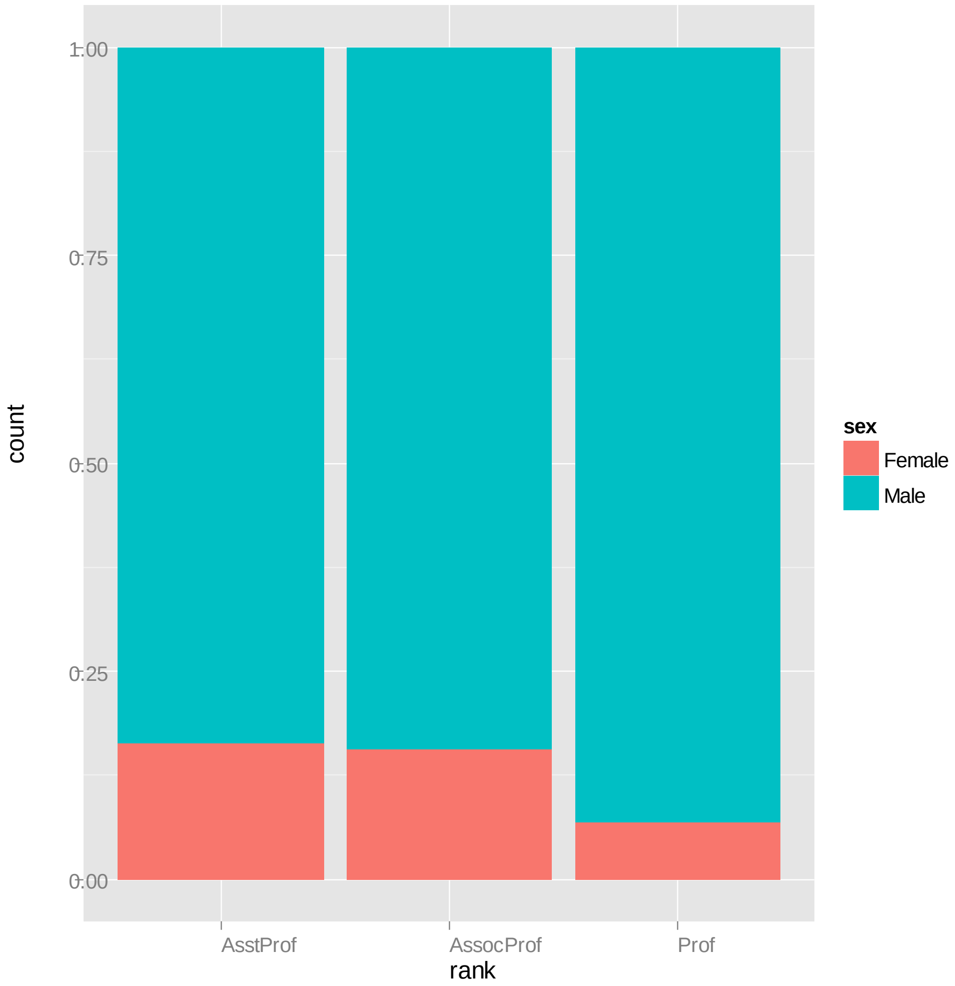
Grouped bar plot

```
ggplot(data=Salaries,
aes(x=rank, fill=sex)) +
geom_bar(
position="dodge")
```



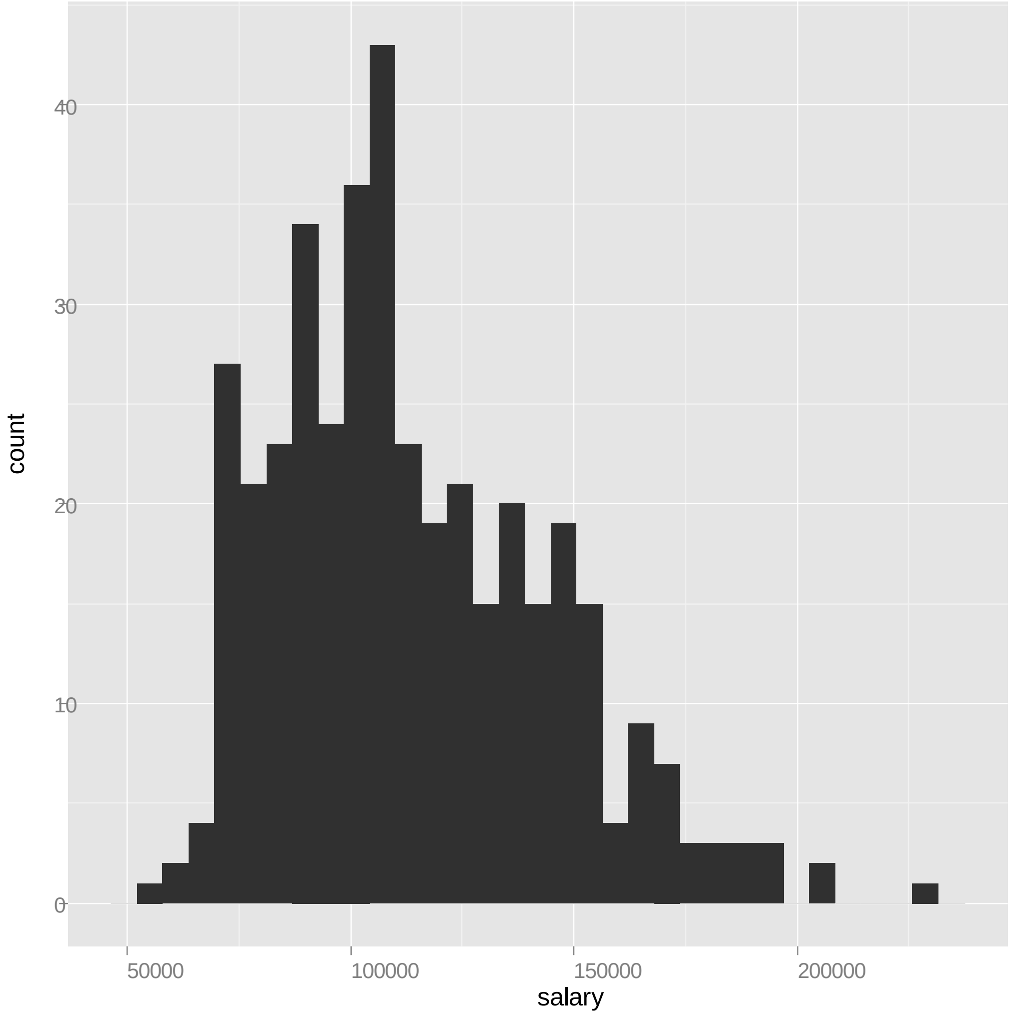
Spinogram

```
ggplot(data=Salaries,
aes(x=rank, fill=sex)) +
geom_bar(
position="fill")
```



Histogram

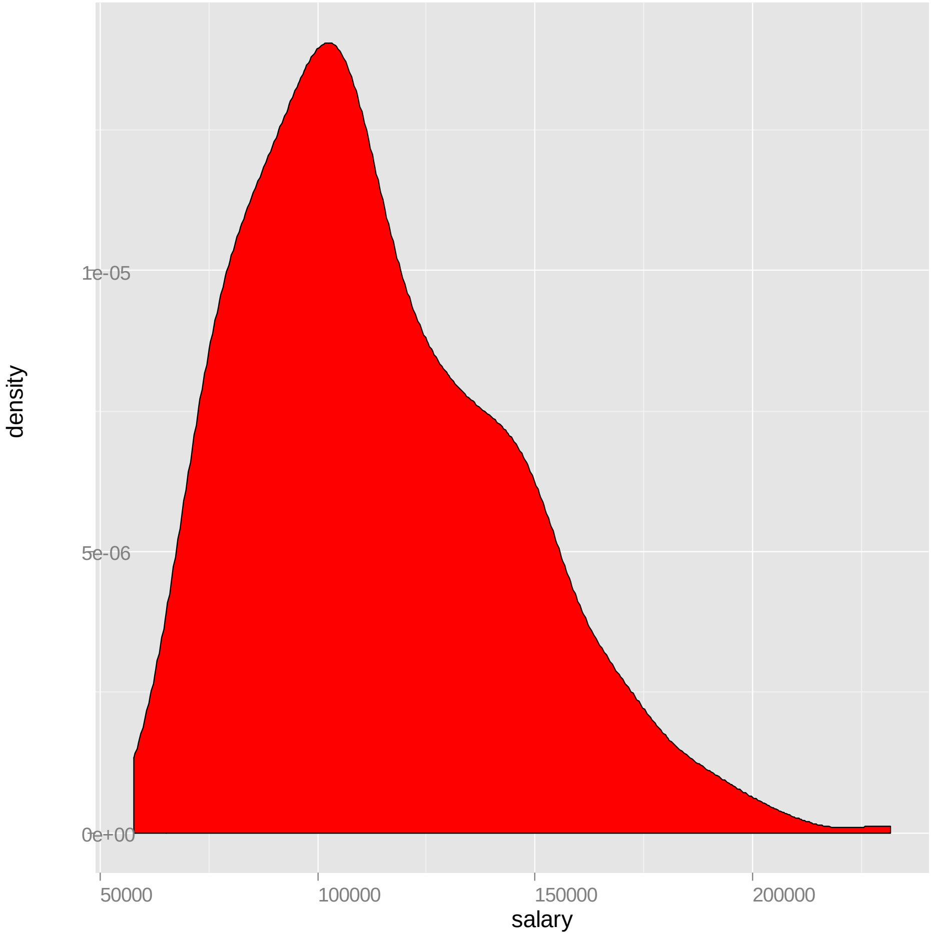
```
ggplot(data=Salaries,
aes(x=salary)) +
geom_histogram()
```



common geom_histogram options:
binwidth
color (border)
fill

Kernel density plot

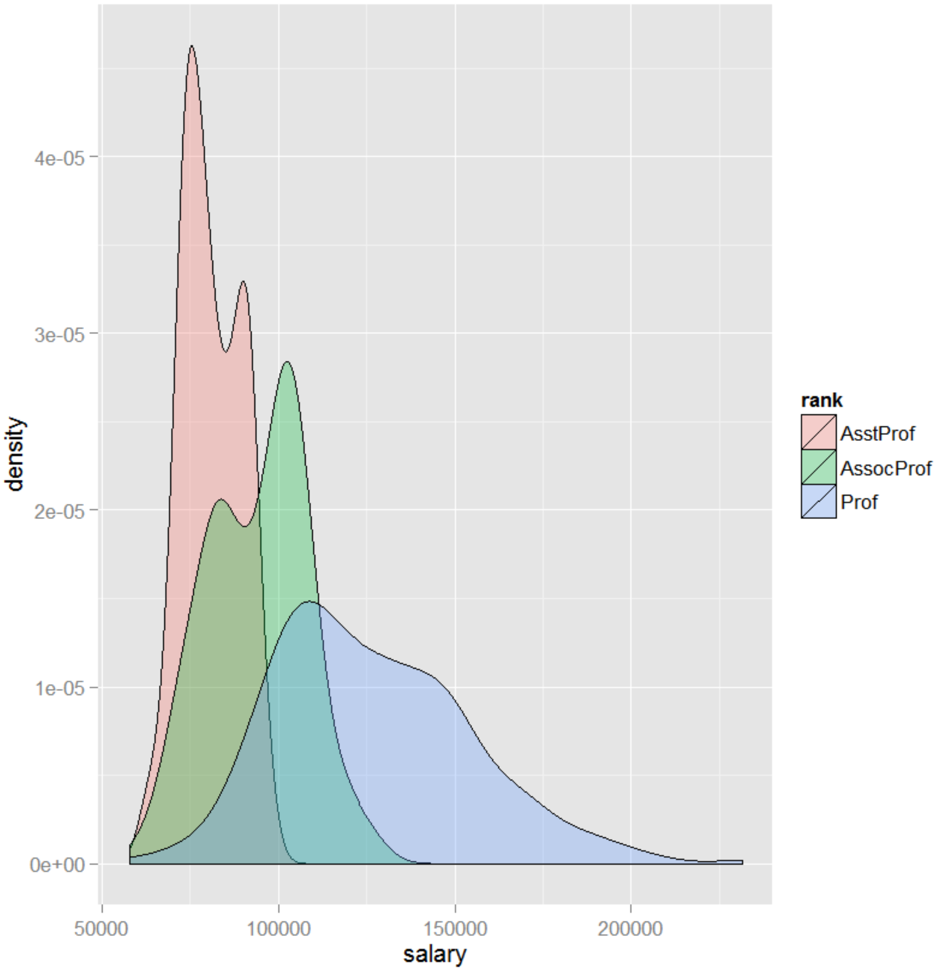
```
ggplot(data=Salaries,
  aes(x=salary)) +
geom_density(fill="red")
```



common geom_density options:
fill
colour
alpha

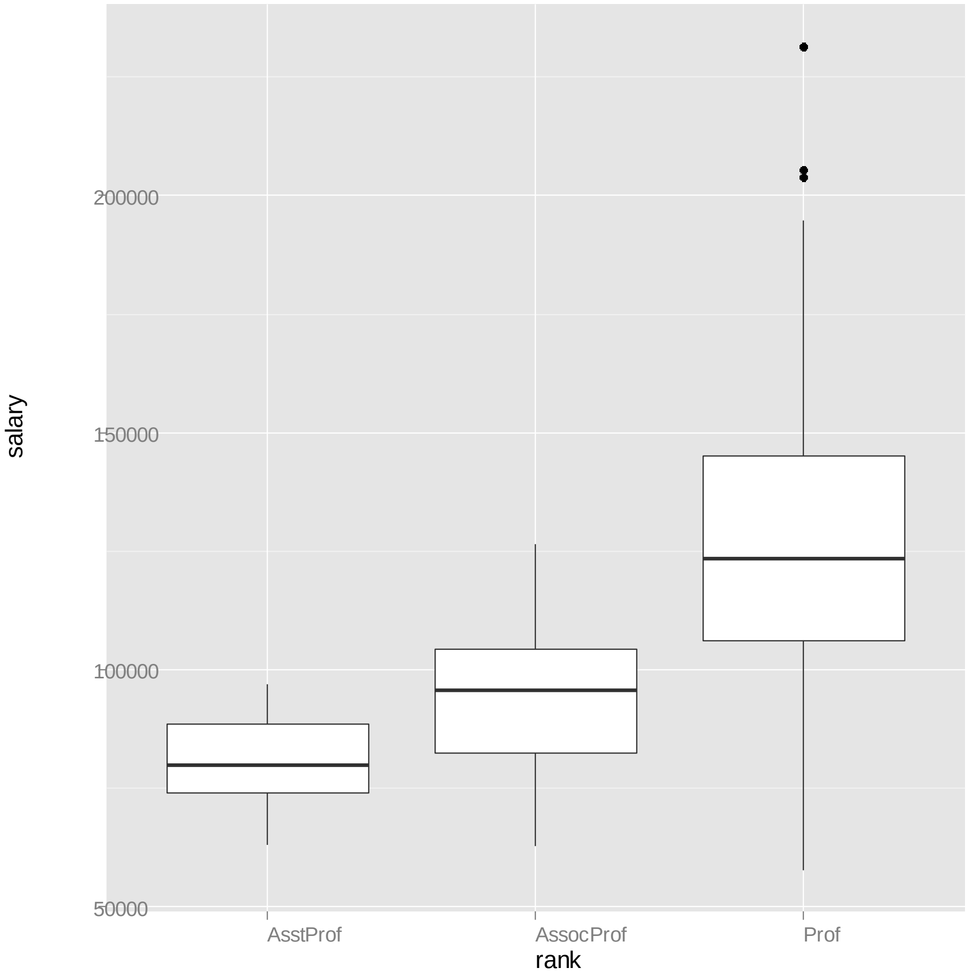
Kernel density plot - multiple groups

```
ggplot(data=Salaries,
aes(x=salary, fill=rank))
geom_density(alpha=.3)
```



Box plot

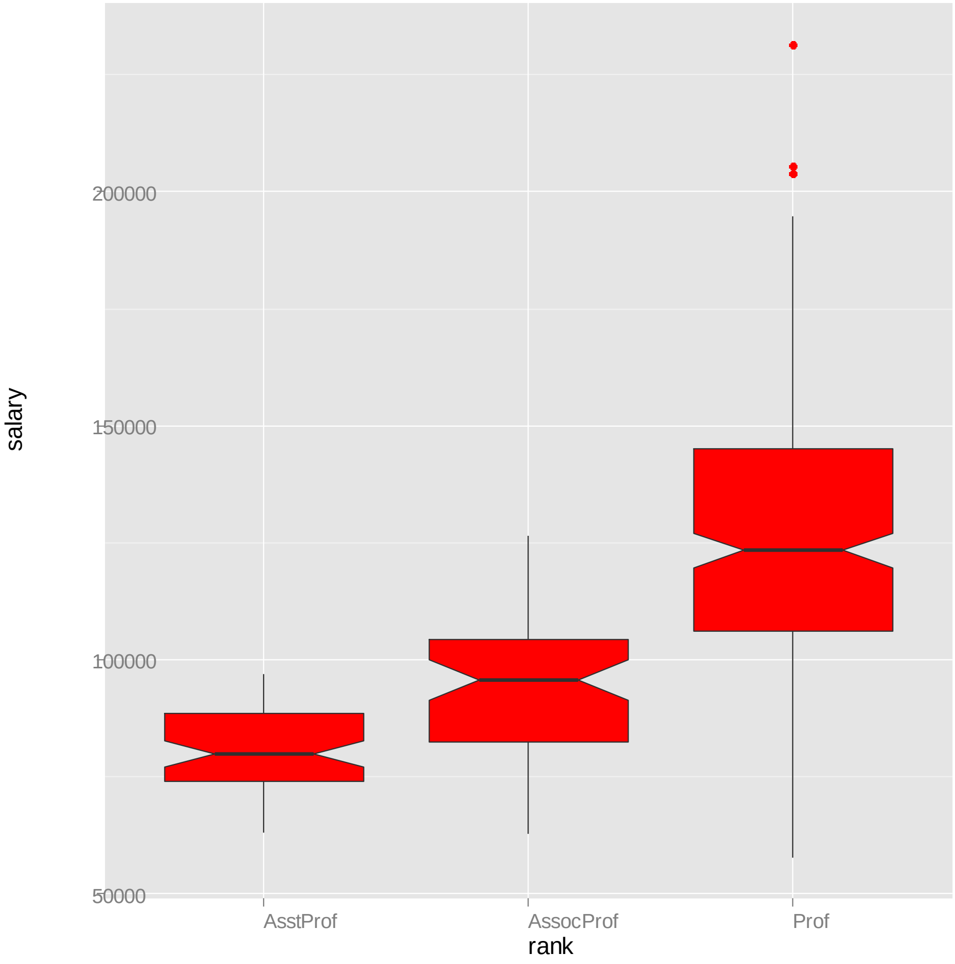
```
ggplot(data=Salaries,
  aes(x=rank, y=salary))
geom_boxplot()
```



common geom_boxplot options:
fill
color
notch (=TRUE or FALSE)
outlier. color shape size

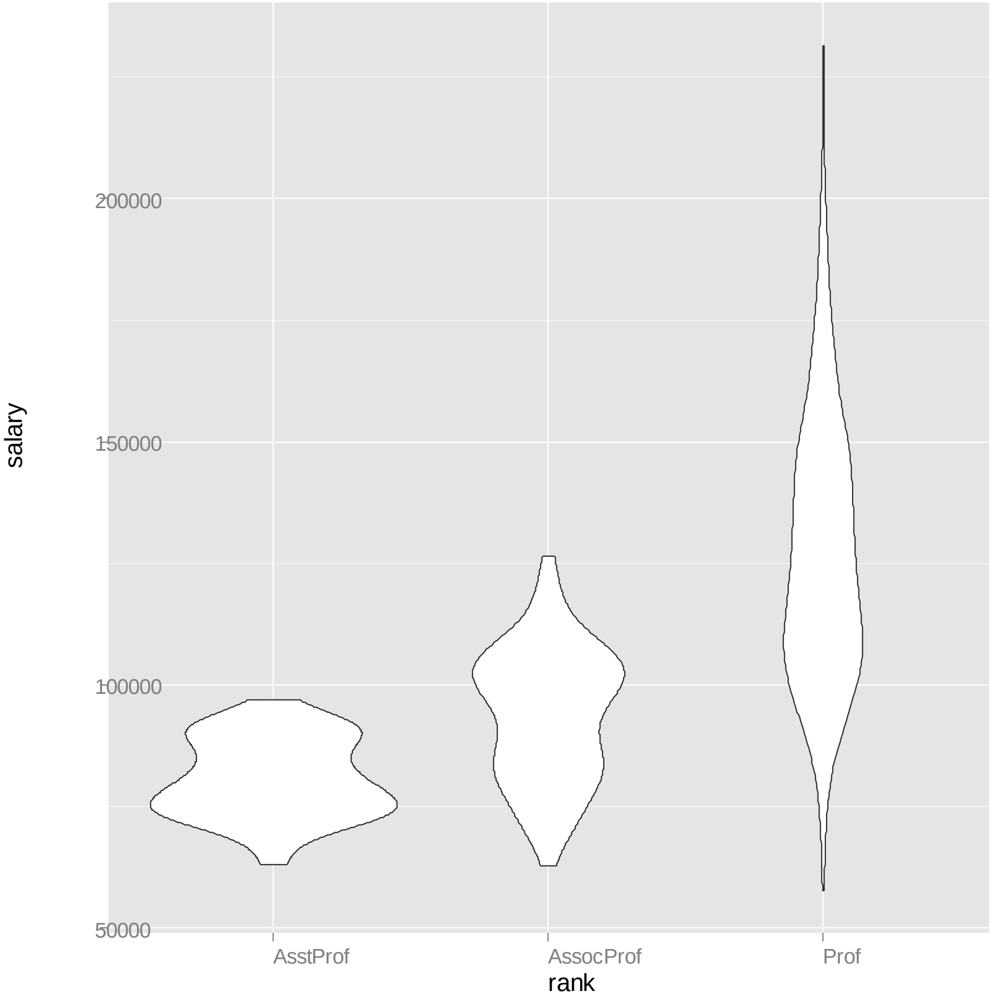
Notched box plot

```
ggplot(data=Salaries,
aes(x=rank, y=salary)) +
geom_boxplot(fill="red",
   notch=TRUE,
   outlier.size=2,
   outlier.color="red")
```



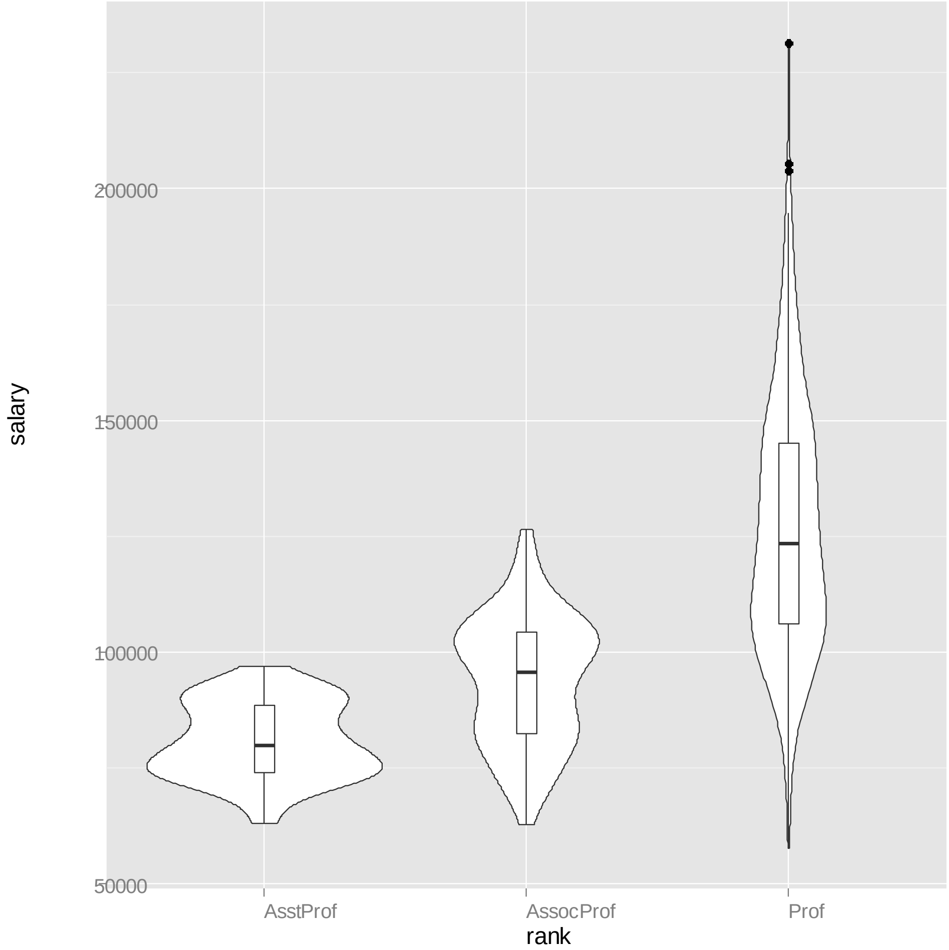
Violin plot

```
ggplot(data=Salaries,
  aes(x=rank, y=salary))
geom_violin()
```



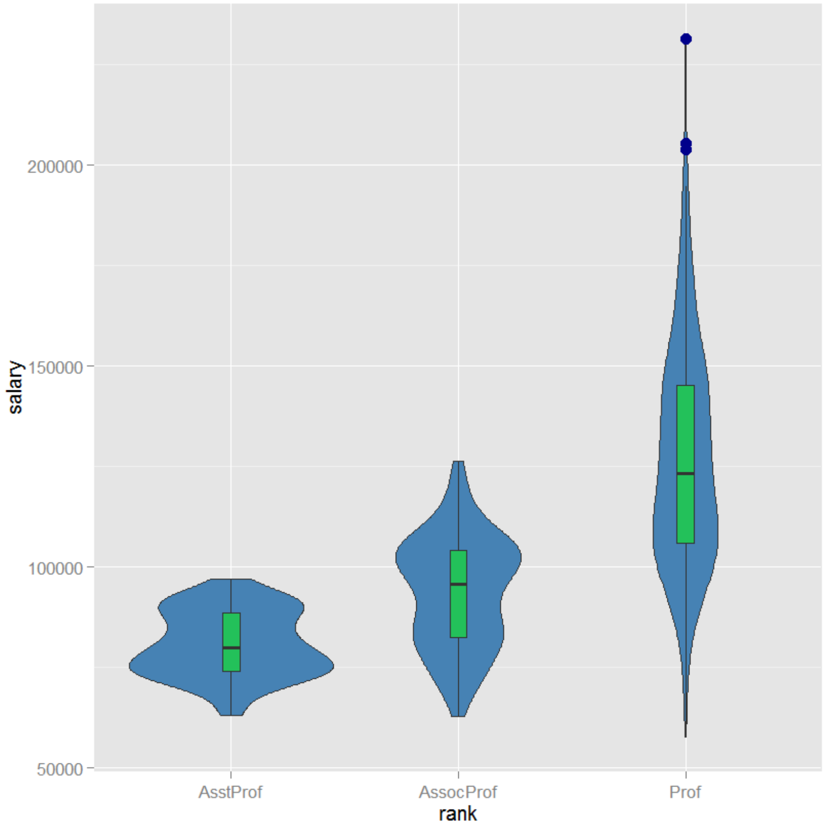
Combining violin and box plots

```
ggplot(data=Salaries,
  aes(x=rank, y=salary)) +
geom_violin() +
geom_boxplot(width=.1)
```



Combining violin and box plots

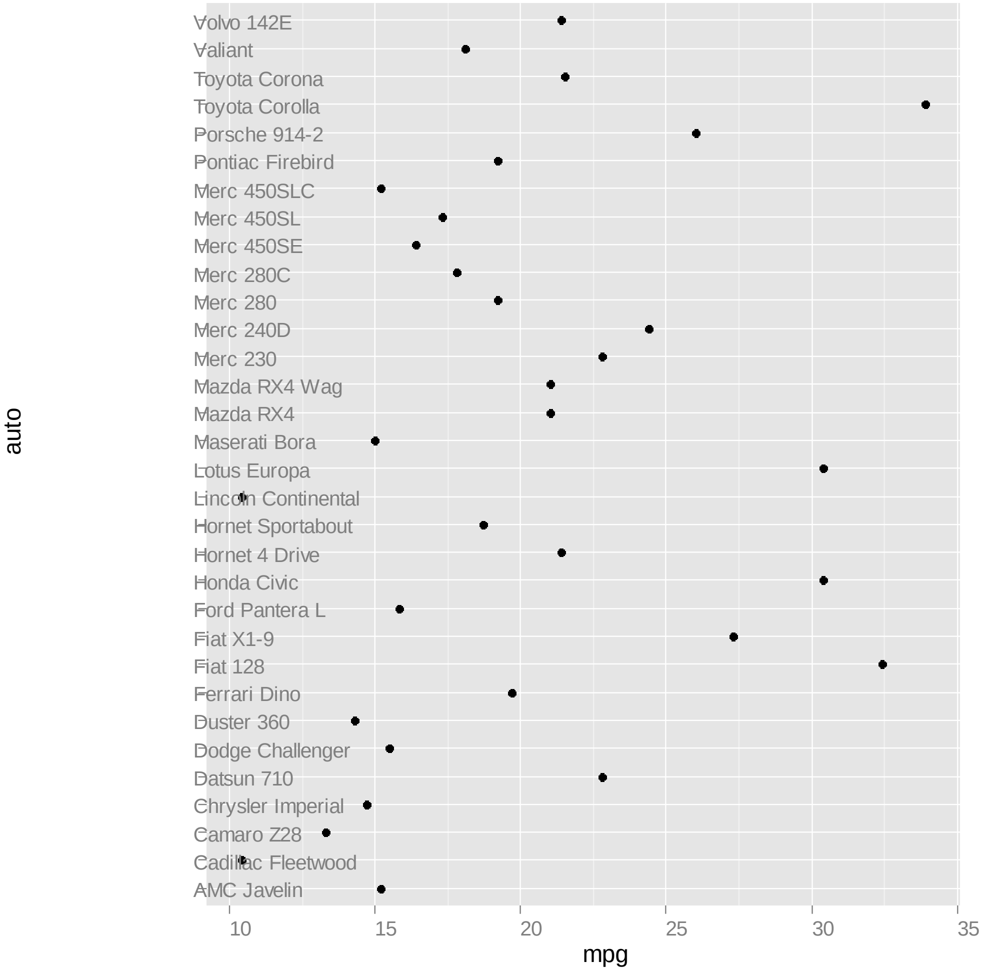
```
ggplot(data=Salaries,
aes(x=rank, y=salary)) +
geom_violin(fill="steelblue") +
geom_boxplot(fill="green",
  alpha=.5,
  width=.1,
  outlier.size=3,
  outlier.colour="darkblue")
```



Dot plot

```
df <- mtcars
df$cars <- row.names(df)

ggplot(data=df,
  aes(x=mpg, y=auto)) +
  geom_point()
```



Sorted Dot plot

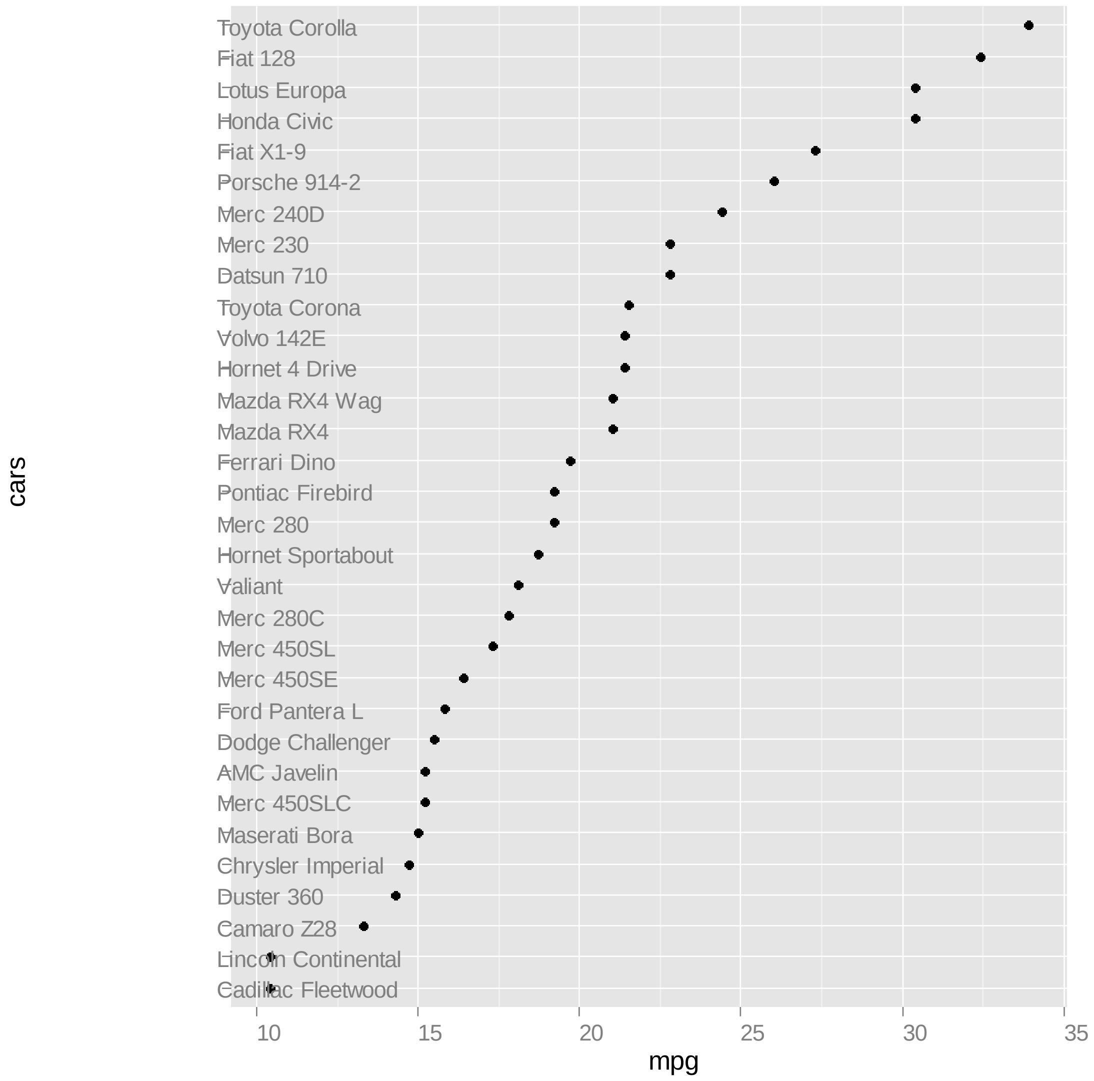
```
df <- mtcars[order(mtcars$mpg),]

levels <- c(1:nrow(df))

df$cars <- factor(levels, labels=row.names(df))

ggplot(data=df, aes(x=mpg, y=cars)) + geom_point()
```

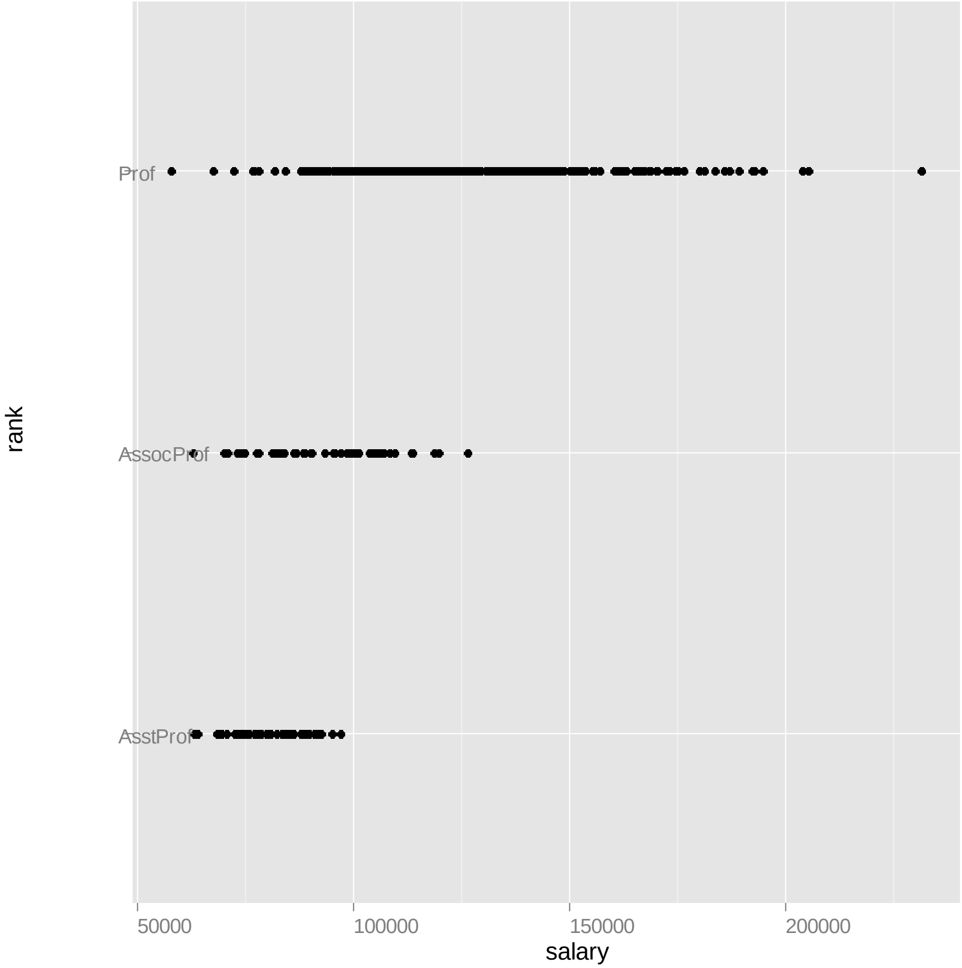
Sorted Dot Plot



Strip plot

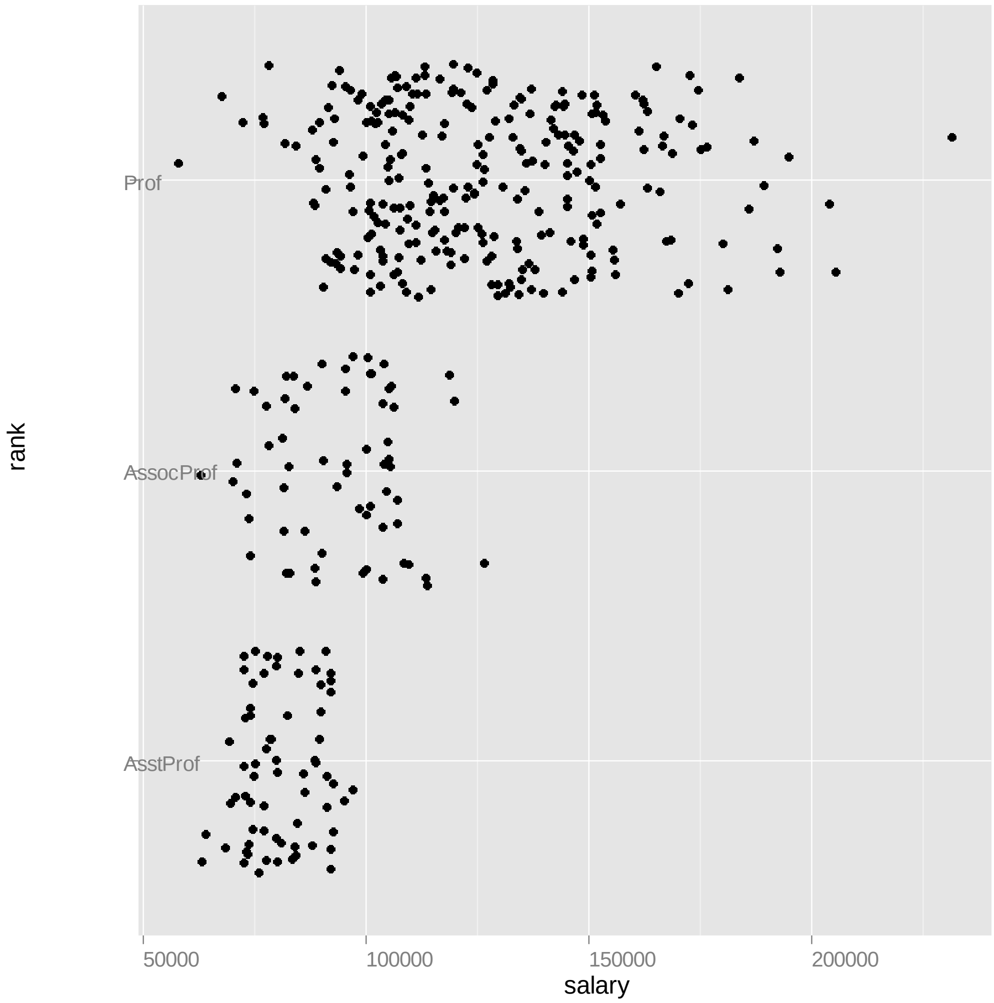
```
ggplot(data=Salaries,
  aes(x=salary, y=rank))
  geom_point()
```

common geom_point options:
color
fill
alpha
shape
size



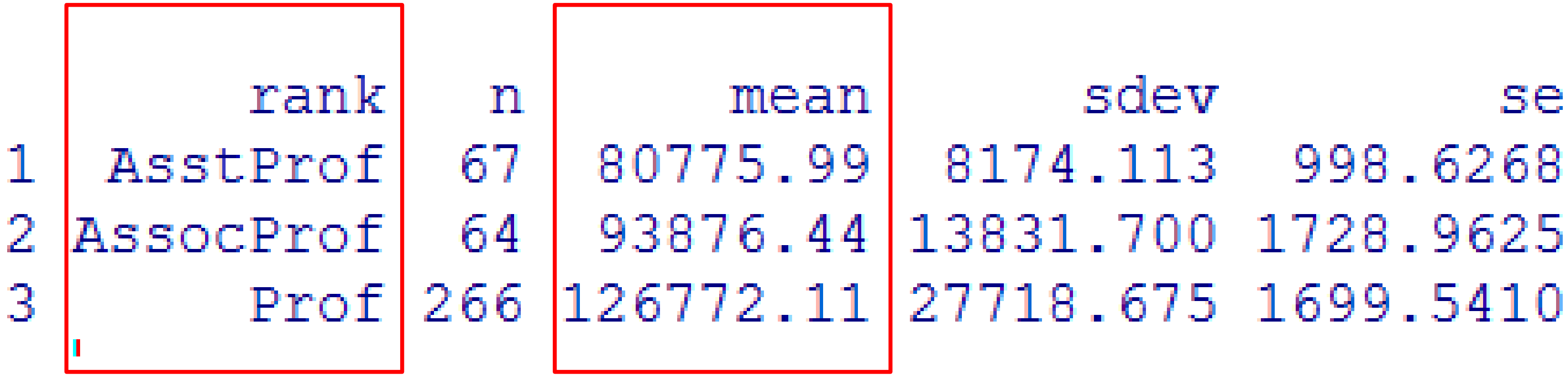
Jittered Strip plot

```
ggplot(data=Salaries,
  aes(x=salary, y=rank))
  geom_jitter()
```



Mean plot

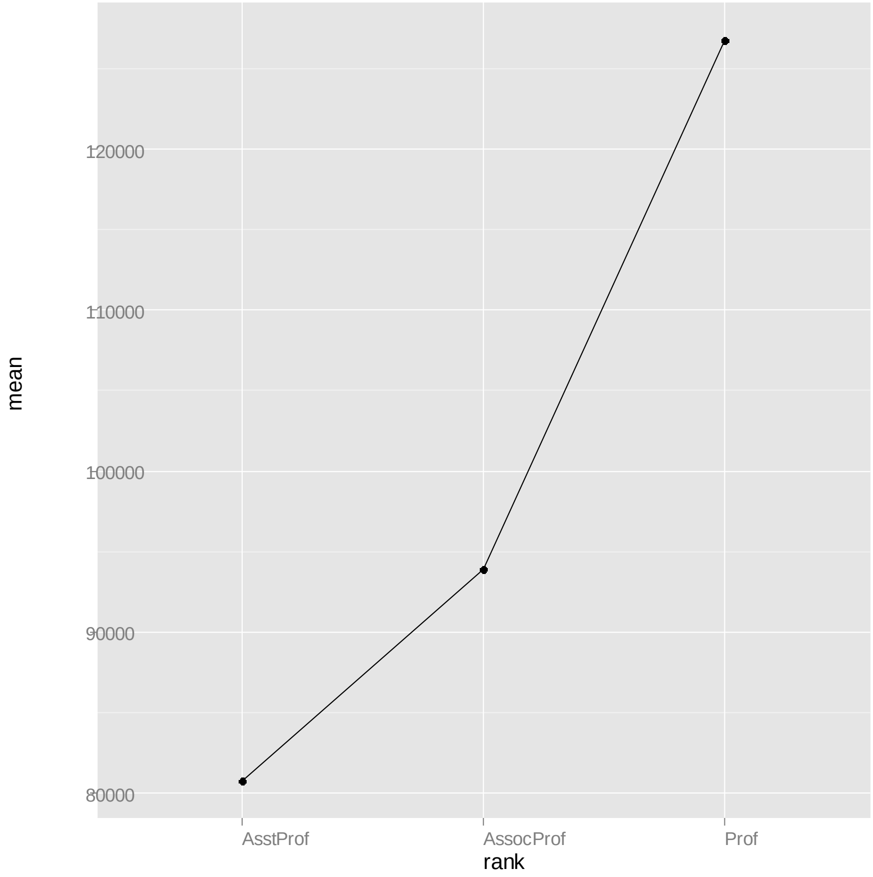
```
library(dplyr)
df <- group_by(Salaries, rank) %>%
      summarise(n=n(), mean=mean(salary),
                se=sd(salary)/sqrt(n))
```



```
       rank   n      mean      sdev        se
1  AsstProf  67  80775.99  8174.113  998.6268
2 AssocProf  64  93876.44 13831.700 1728.9625
3      Prof 266 126772.11 27718.675 1699.5410
```

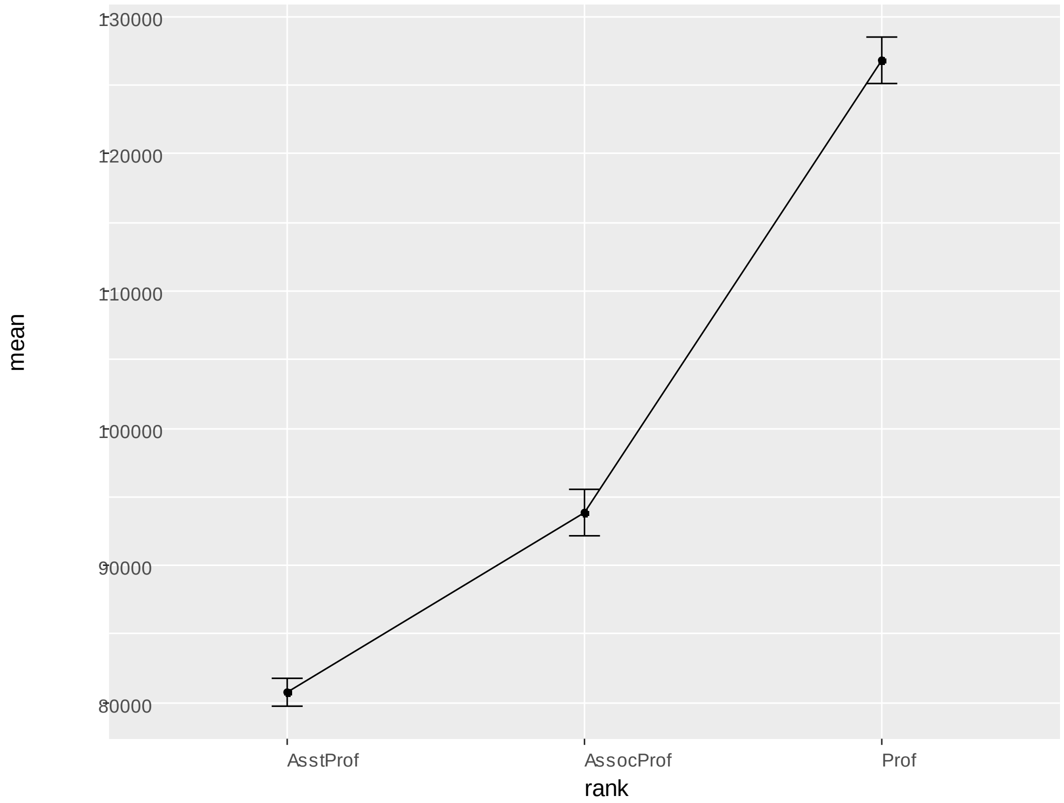
Mean plot

```
ggplot(df, aes(x=rank, y=mean, group=1)) +
  geom_line() +
  geom_point()
```



Mean plot with standard errors

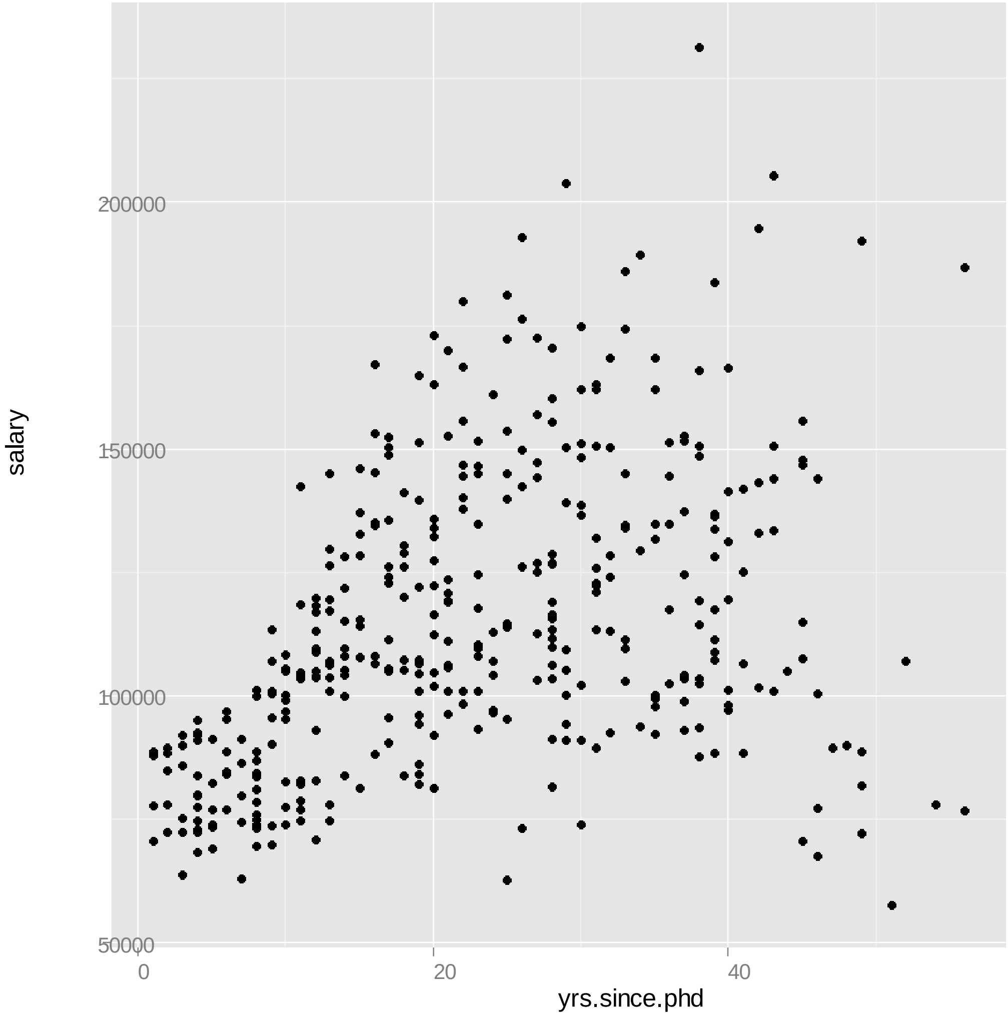
```
ggplot(df, aes(x=rank, y=mean, group=1)) +
  geom_errorbar(aes(
    ymin=mean-se,
    ymax=mean+se,
    width=.1)) +
  geom_line() +
  geom_point()
```



Scatter plot

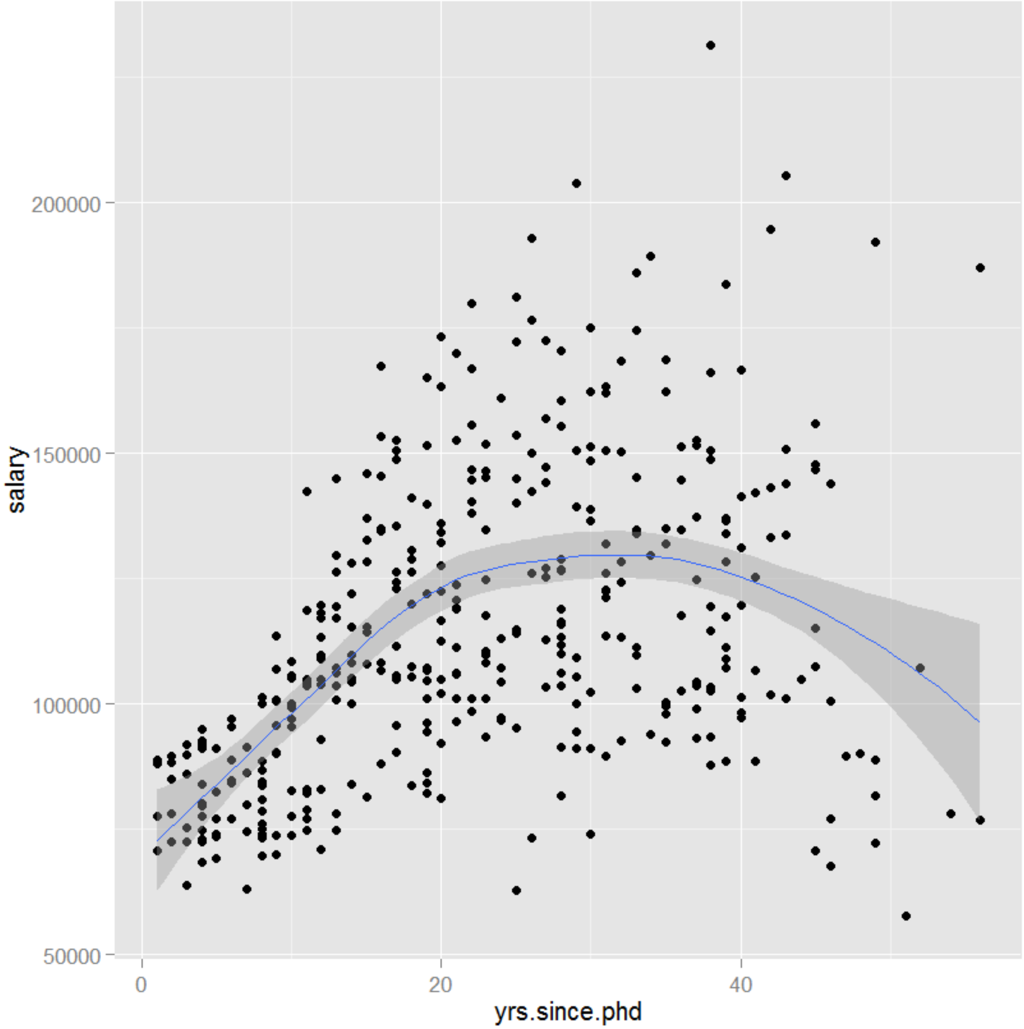
```
ggplot(data=Salaries,
  aes(x=yrs.since.phd,
      y=salary)) +
geom_point()
```

common geom_point options:
color
fill
alpha
shape
size



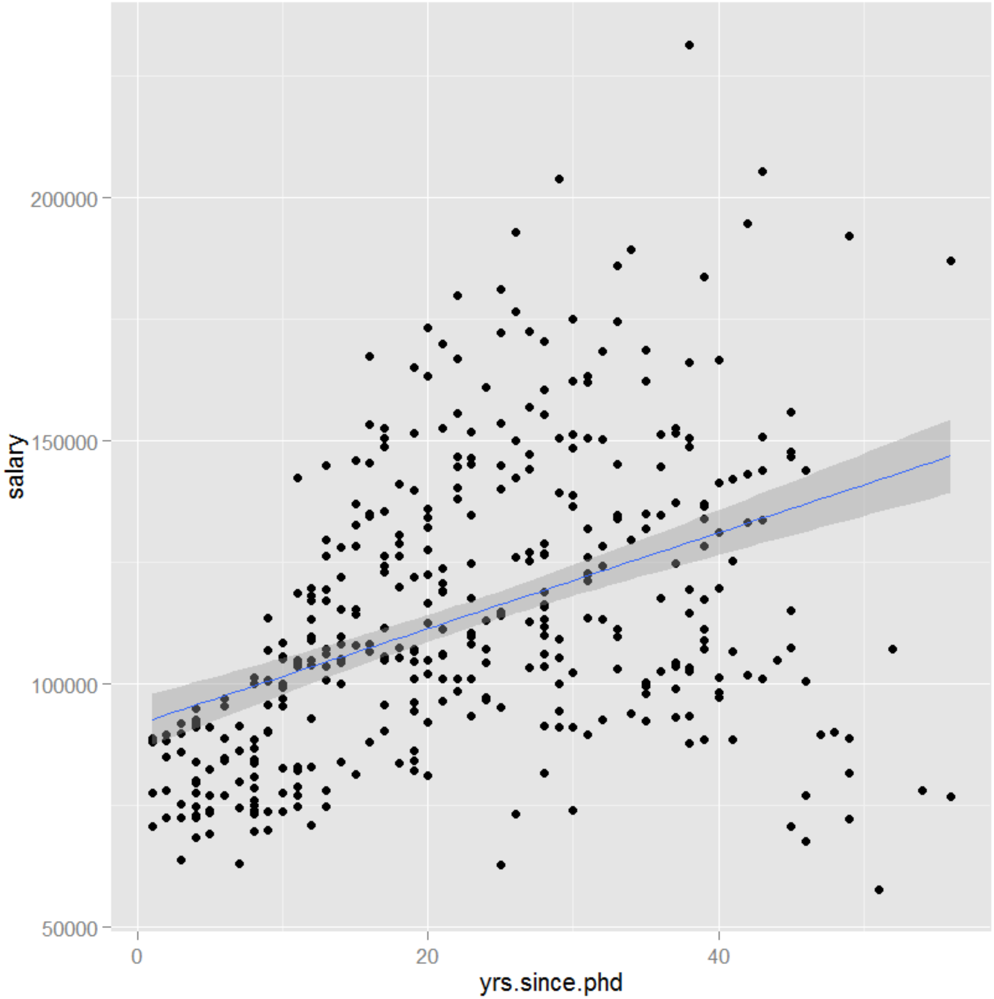
Scatterplot with fit

```
ggplot(data=Salaries,
   aes(x=yrs.since.phd,
       y=salary)) +
geom_point() +
geom_smooth()
```



Scatterplot with fit

```
ggplot(data=Salaries,
  aes(x=yrs.since.phd,
      y=salary)) +
geom_point() +
geom_smooth(method="lm",
            formula=y~x)
```



Grouping

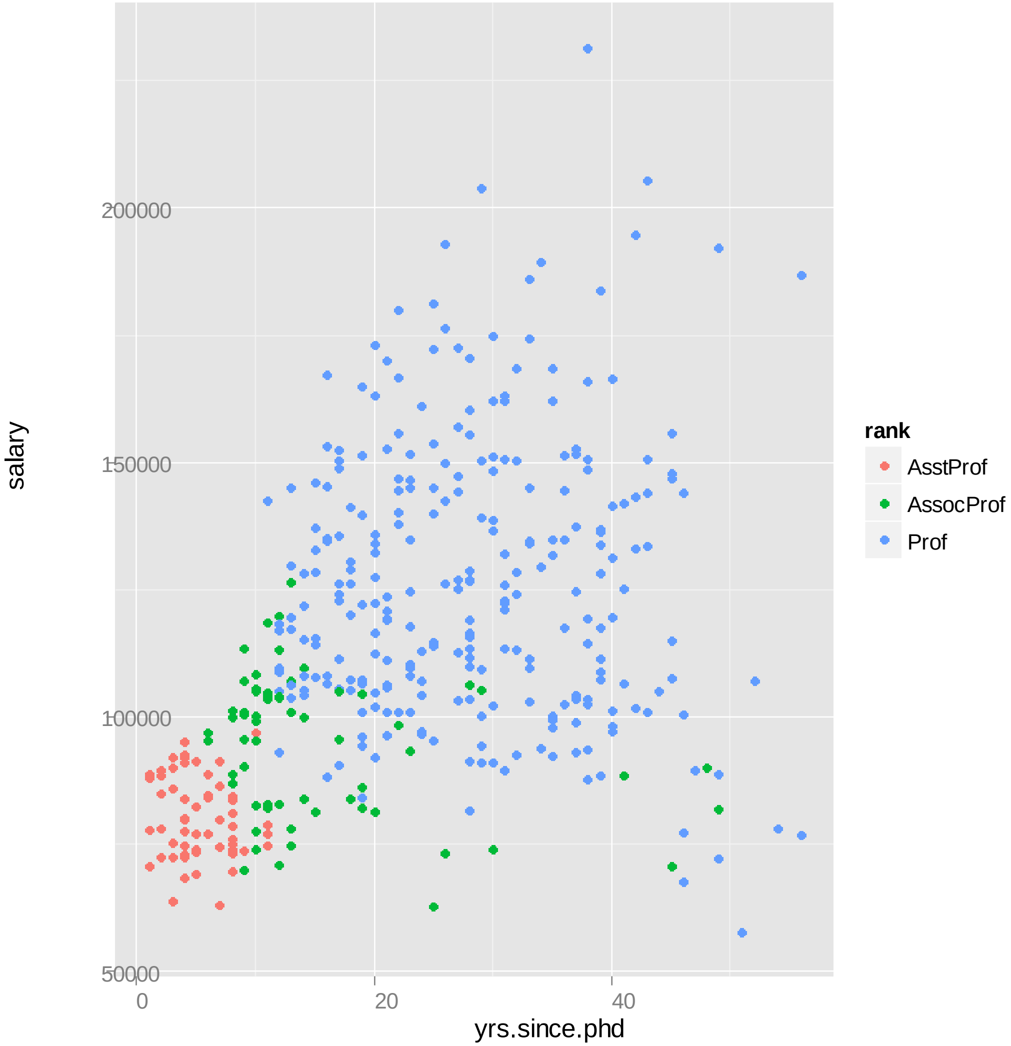
Add

- *color,*
- *shape,*
- *size,*
- *alpha*

to aes or the geom_xxx()

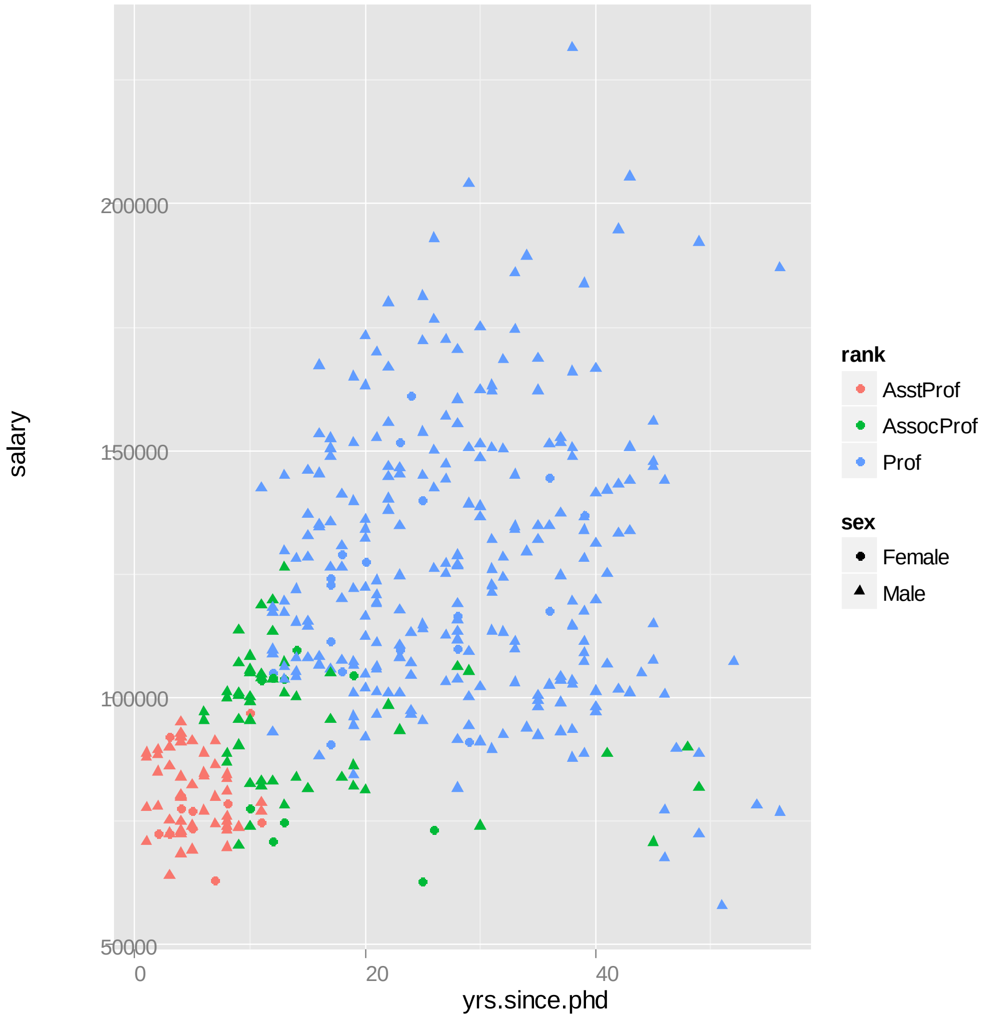
Grouping

```
ggplot(data=Salaries,
  aes(x=yrs.since.phd,
      y=salary,
      color=rank)) +
geom_point()
```



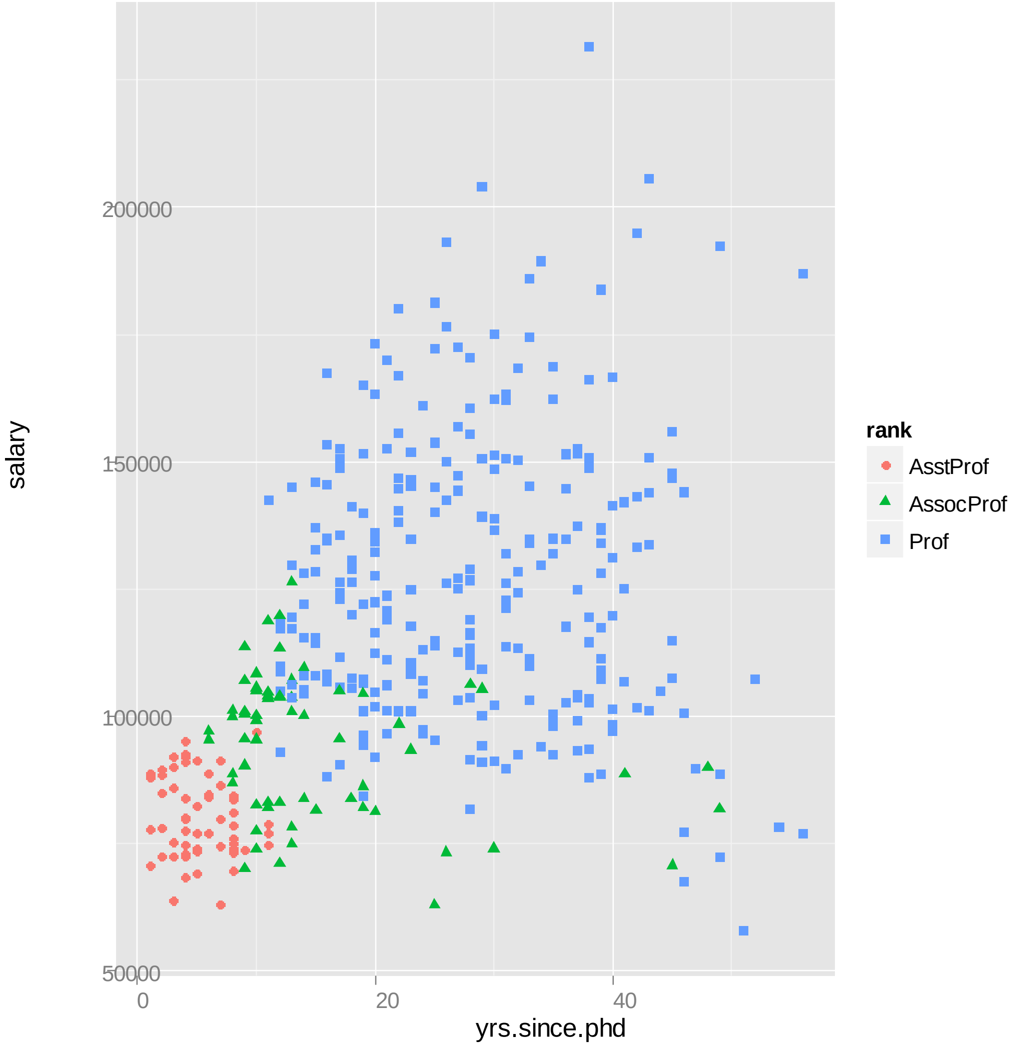
Grouping

```
ggplot(data=Salaries,
  aes(x=yrs.since.phd,
      y=salary,
      color=rank,
      shape=sex)) +
geom_point()
```



Grouping

```
ggplot(data=Salaries,
  aes(x=yrs.since.phd,
      y=salary,
      color=rank,
      shape=rank)) +
geom_point()
```



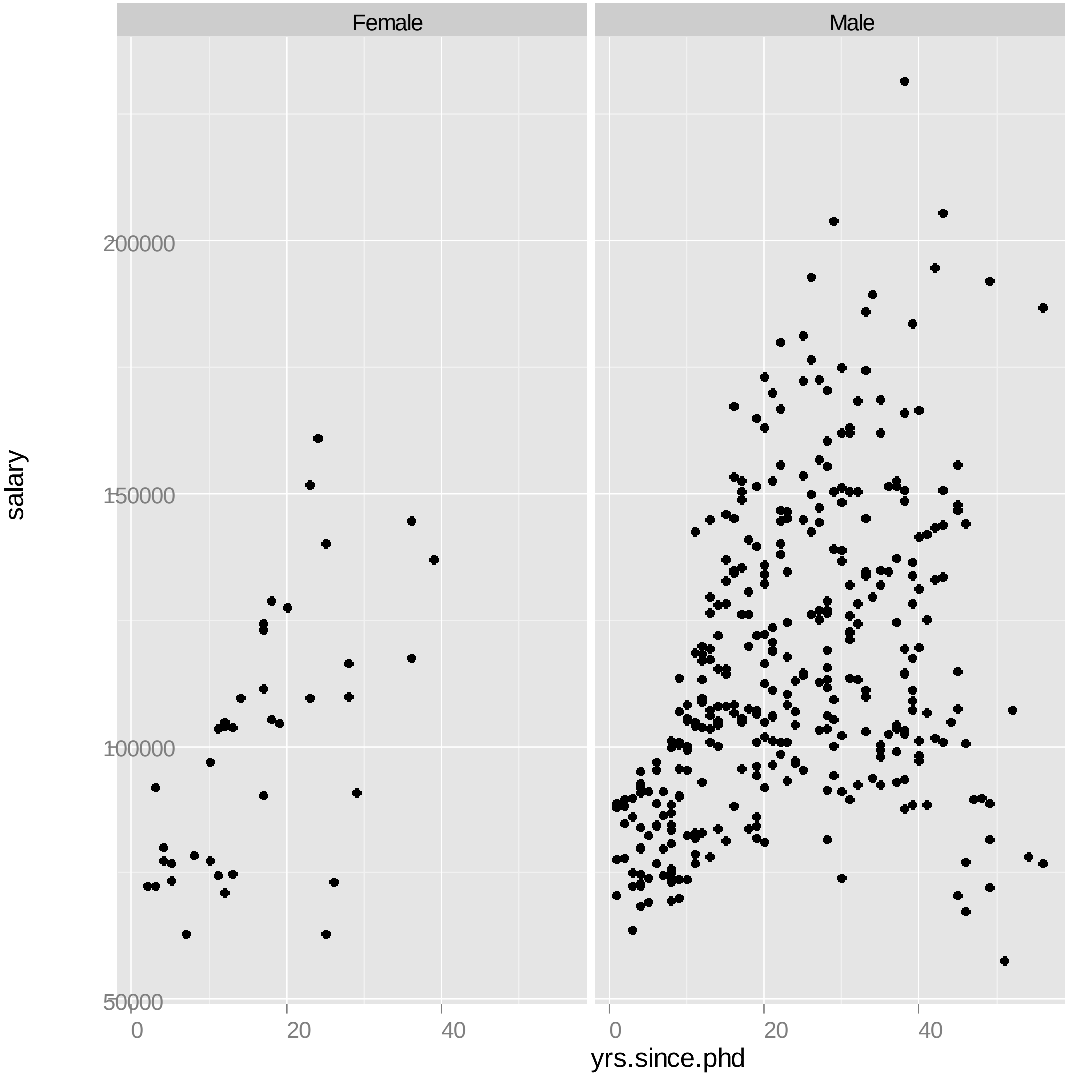
Facets

```
facets_grid( rowvar ~ colvar)
```

`facets_grid( . ~ colvar)`	`just columns`
`facets_grid(rowvar ~ .)`	just rows
`facets_wrap(~ var, ncol=#)`	one classification variable wrapped to fill page

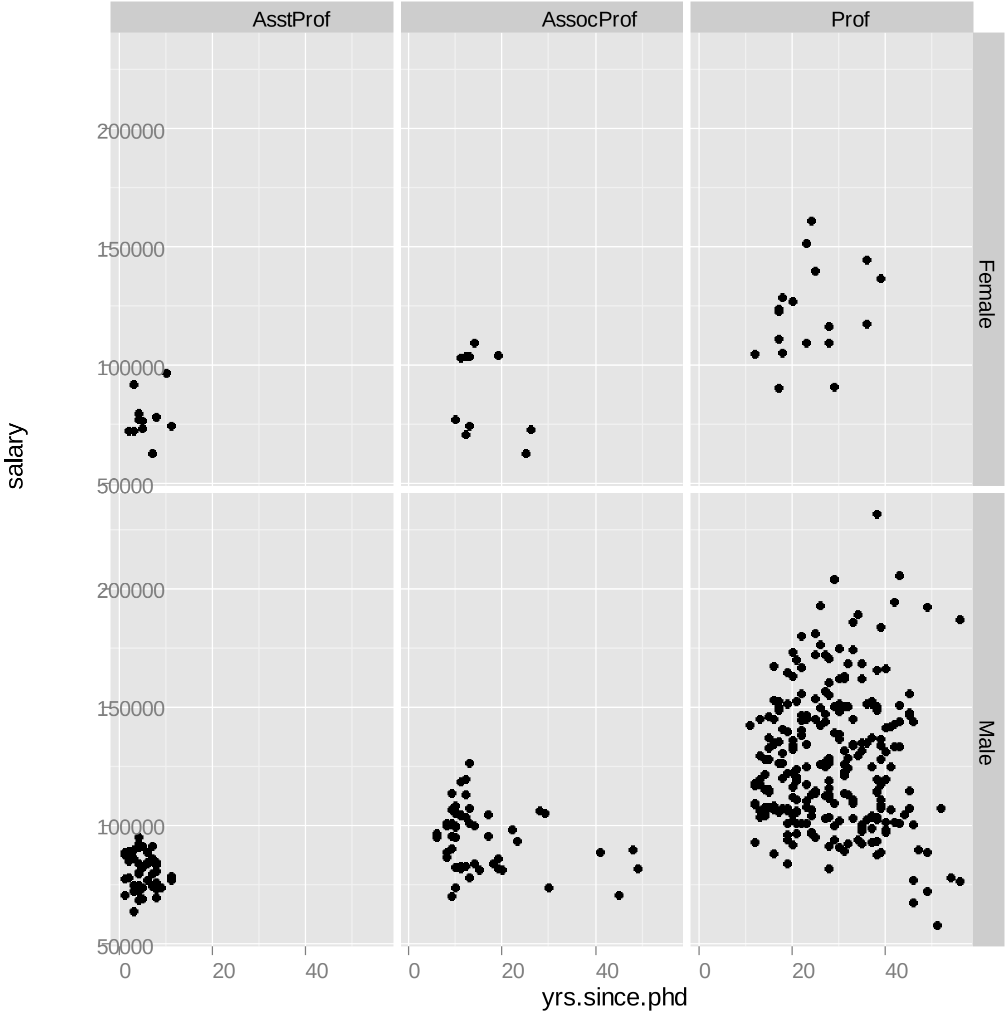
Facets

```
ggplot(data=Salaries,
  aes(x=yrs.since.phd,
      y=salary)) +
geom_point() +
facet_grid(. ~ sex)
```



Facets

```
ggplot(data=Salaries,
  aes(x=yrs.since.phd,
      y=salary)) +
geom_point() +
facet_grid(sex ~ rank)
```



Saving your work

- `ggsave(filename="filename.ext", plot=)`
  - ext can be

    `eps, ps, tex, pdf, jpeg, tiff, png, bmp, svg, wmf`
  - plot defaults to last one created
  - wmf on windows platforms only
  - svg can be edited using Inkscape

Learning more

- Hadley Wickham – http://docs.ggplot2.org/
- Winston Chang- http://wiki.stdout.org/rcookbook/Graphs/