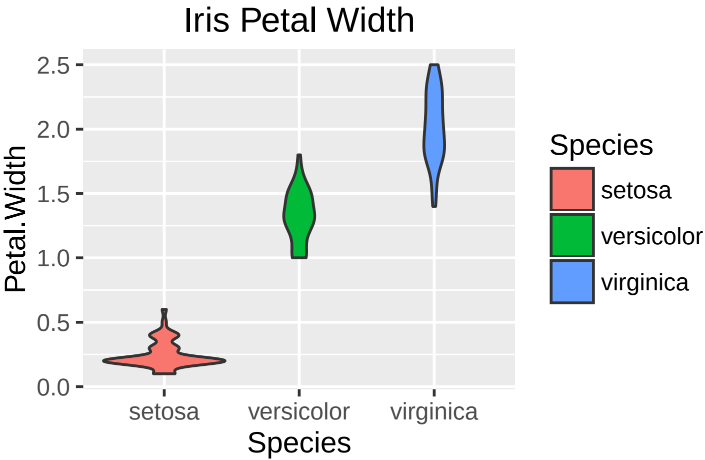
# Introduction to ggplot

Lingge Li

2/21/2017

# Why ggplot

- Beautiful aesthetics
- Flexible and powerful



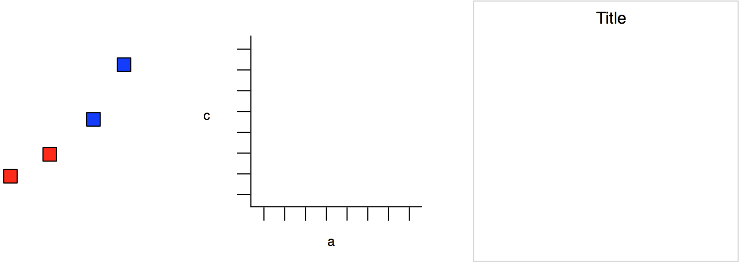
# However. . .

- Syntax slightly compilcated at first glance

```
ggplot(data=iris, aes(x=Species, y=Petal.Width)) +
  geom_violin(aes(fill=Species)) +
  labs(title='Iris Petal Width')
```

# Layered grammar of graphics

- ggplot2 follows a specific grammar of graphics



**Geoms**

**Guides**

(from scales and coordinate systems)

**Plot**

*Example taken from Hadley Wickham's book*
`http://vita.had.co.nz/papers/layered-grammar.pdf`

# How to make a plot

- Geometric objects (geom)
- Aesthetic mapping (aes)
- Statistical transformation (stat)
- Scales and coordinate system

# Geoms

- Wide range of geometric objects from points to complex shapes
- geom_point, geom_line, geom_histogram, geom_boxplot. . .
- Multiple geometric objects on the same plot with +

# Aesthetics

- Coordinate positions (always needed)
- Colour, fill, shape, size. . .

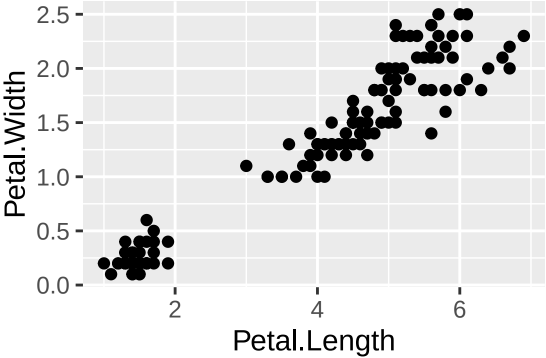
# Data + mapping

- aes() maps a dataframe to geom
- Each geom can have its own mapping

```
geom_point(data, aes(x, y))
```

# Scatterplot

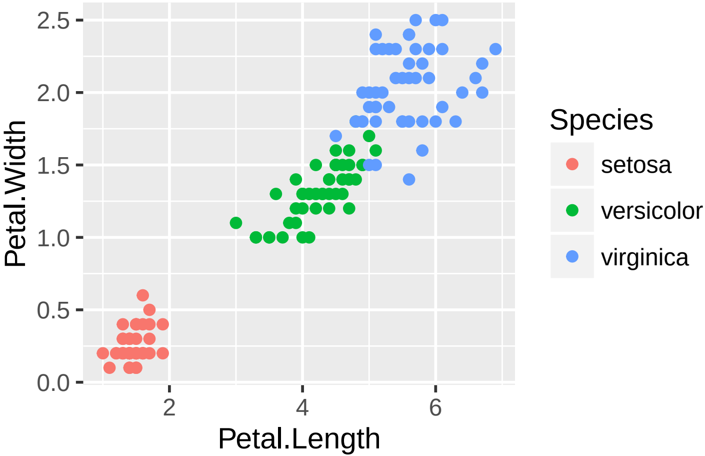
```
ggplot() + geom_point(data=iris,
                   aes(x=Petal.Length, y=Petal.Width))
```



- The code below equivalent and more commonly seen

```
ggplot(data=iris, aes(x=Petal.Length, y=Petal.Width)) +
  geom_point()
```

# Colour and shape

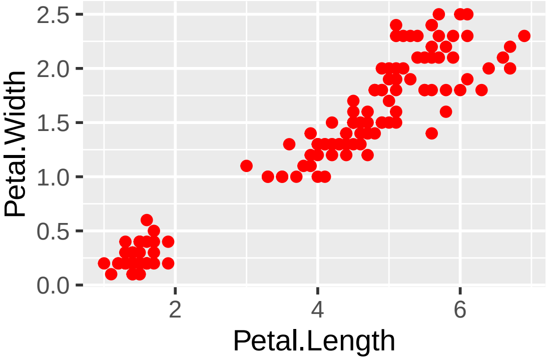


How would you change the shape?

# Mapping vs setting

- What is the key difference

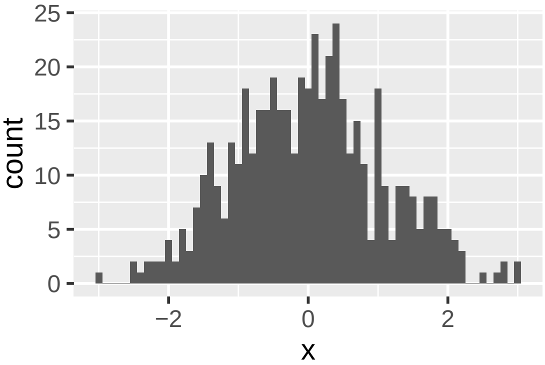
```
ggplot(data=iris, aes(x=Petal.Length, y=Petal.Width)) +
  geom_point(colour='red')
```



# Stat

- Draw a histogram the hard way

```
x <- rnorm(500)
temp <- as.data.frame(x)
ggplot(temp, aes(x=x)) +
  geom_histogram(stat='bin', binwidth=0.1)
```

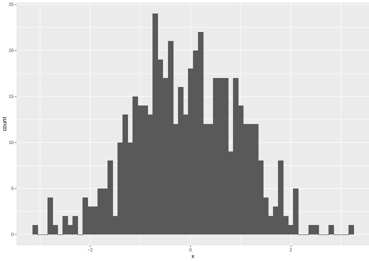


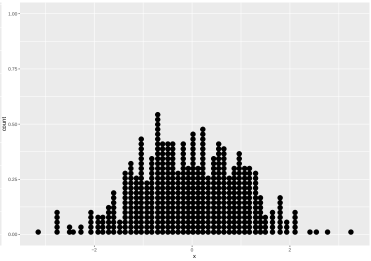
- Most plots do not require stat

# Quick plot

- qplot similar to base plot
- http://docs.ggplot2.org/current/qplot.html

```r
x <- rnorm(500)
qplot(x, geom='histogram', binwidth=0.1)
qplot(x, geom='dotplot', binwidth=0.1)
```



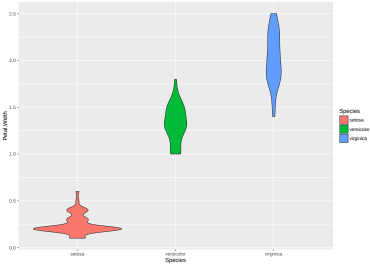


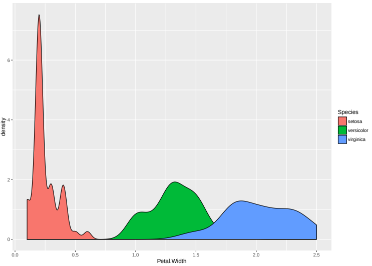
# Fancy plots

- Basic plots done fancily with ggplot

```
ggplot(data=iris, aes(x=Species, y=Petal.Width)) +
  geom_violin(aes(fill=Species))

ggplot(data=iris, aes(x=Petal.Width)) +
  geom_density(aes(fill=Species))
```

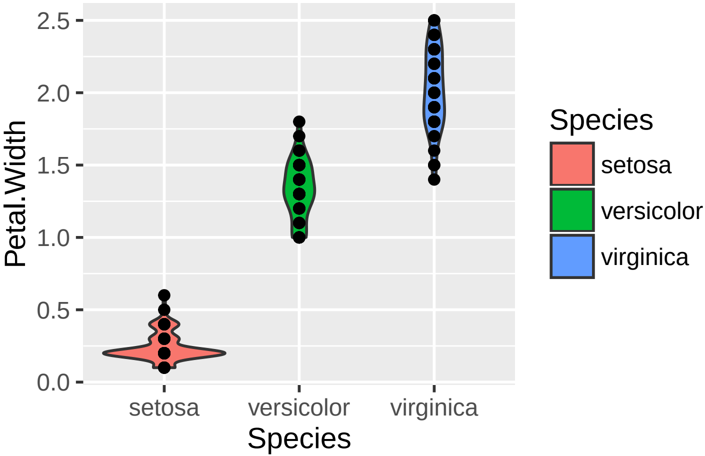




# Violin plot with points

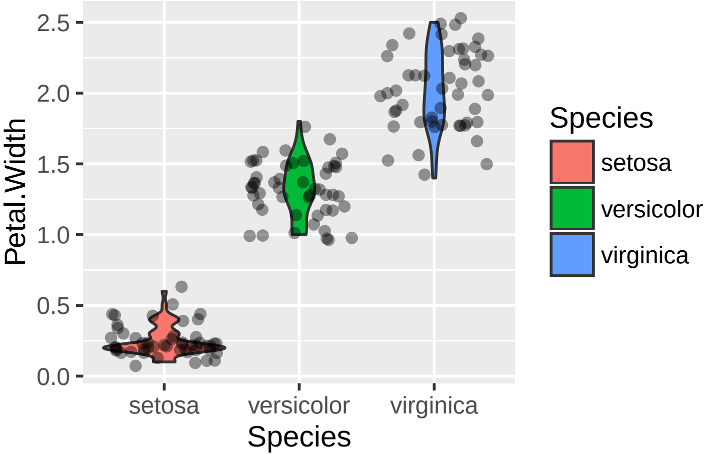
- geom_violin + geom_point

```
ggplot(data=iris, aes(x=Species, y=Petal.Width)) +
  geom_violin(aes(fill=Species)) + geom_point()
```



# Jitter

```
ggplot(data=iris, aes(x=Species, y=Petal.Width)) +
  geom_violin(aes(fill=Species)) +
  geom_jitter(alpha=0.4)
```



# IMDB data

- Dataset contains over 5000 movies and 28 variables from IMDB

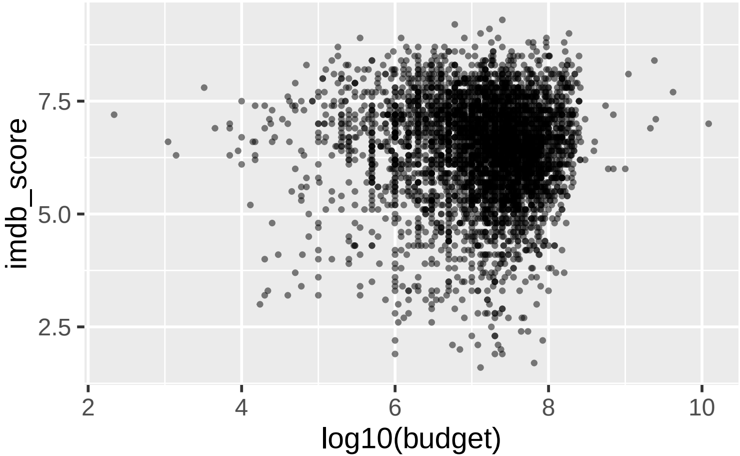
https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

```
imdb <- read.csv("~/Downloads/movie_metadata.csv")
colnames(imdb)[1:20]
```

```
##  [1] "color"                     "director_name"
##  [3] "num_critic_for_reviews"    "duration"
##  [5] "director_facebook_likes"   "actor_3_facebook_likes
##  [7] "actor_2_name"              "actor_1_facebook_likes
##  [9] "gross"                     "genres"
## [11] "actor_1_name"              "movie_title"
## [13] "num_voted_users"           "cast_total_facebook_li
## [15] "actor_3_name"              "facenumber_in_poster"
## [17] "plot_keywords"             "movie_imdb_link"
## [19] "num_user_for_reviews"      "language"
```

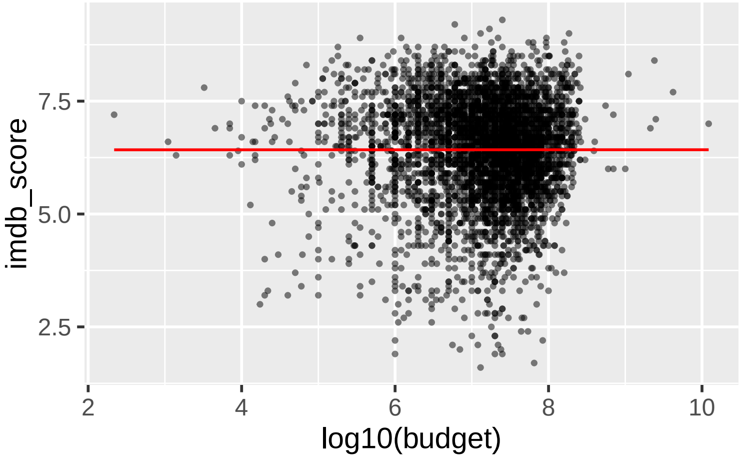
# IMDB score vs log10(budget)

```
imdb <- imdb[!is.na(imdb$imdb_score) & !is.na(imdb$budget),
```



# Add regression line

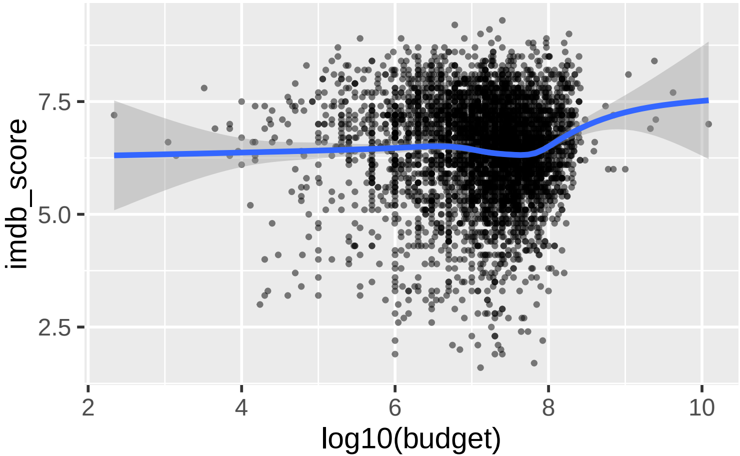
```
imdb$regression <- lm(imdb_score ~ log10(budget), data=imdb
```



# geom_smooth

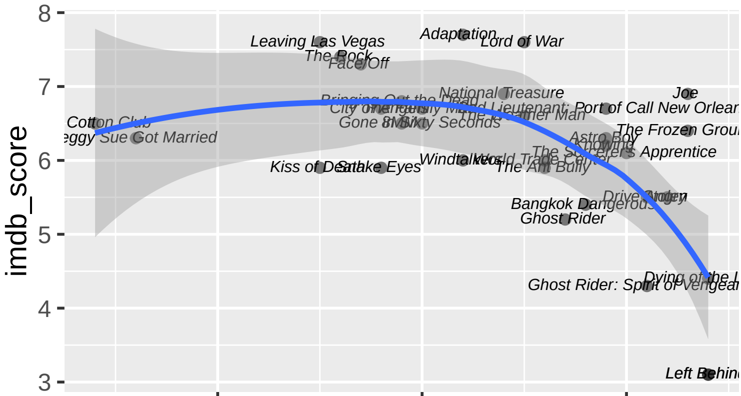
- Kernel smoother

```
ggplot(data=imdb, aes(x=log10(budget), y=imdb_score)) +
  geom_point(alpha=0.5, size=0.5) + geom_smooth()
```

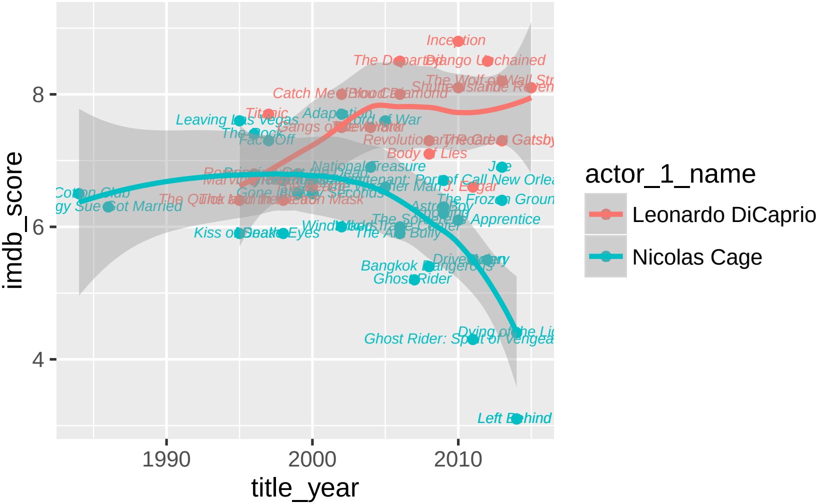


## Nicolas Cage Movies

```
cage <- imdb[imdb$actor_1_name == 'Nicolas Cage', ]
cage$actor_1_name <- as.character(cage$actor_1_name)
ggplot(data=cage, aes(x=title_year, y=imdb_score, label=mov
  geom_point(alpha=0.5) +
  geom_text(fontface='italic', size=2, vjust=1, nudge_y=0.1
  geom_smooth()
```



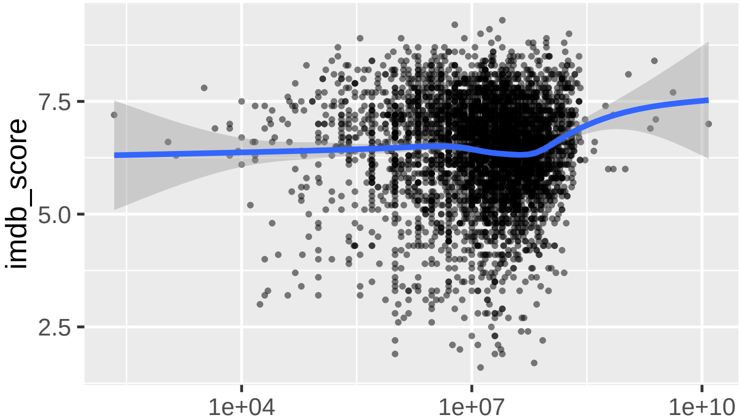
# Compare with Leonardo DiCaprio



# Scale

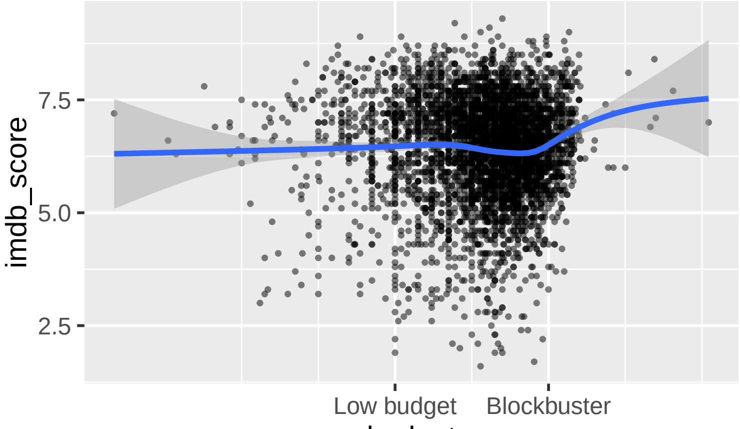
- Transform the scale instead of data
- How is this different from before

```r
ggplot(data=imdb, aes(x=budget, y=imdb_score)) +
  scale_x_log10() +
  geom_point(alpha=0.5, size=0.5) + geom_smooth()
```



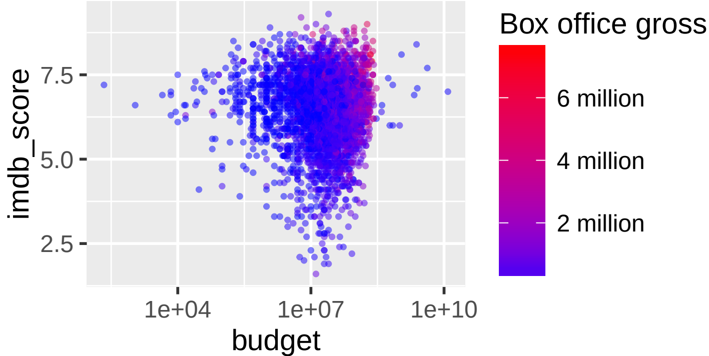
# Custom tick marks and labels

```r
ggplot(data=imdb, aes(x=budget, y=imdb_score)) +
  geom_point(alpha=0.5, size=0.5) + geom_smooth() +
  scale_x_continuous(breaks=c(1e6, 1e8),
                     labels=c('Low budget', 'Blockbuster'),
                     trans='log10')
```

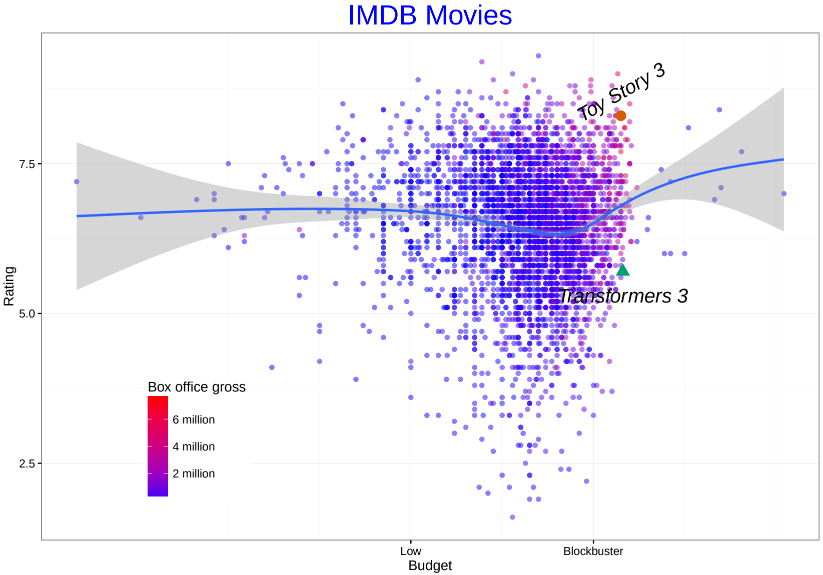


# Colour scale

```
imdb <- imdb[!is.na(imdb$gross), ]
ggplot(data=imdb, aes(x=budget, y=imdb_score)) +
  geom_point(aes(colour=gross), alpha=0.5, size=0.5) +
  scale_x_log10() +
  scale_color_continuous(name='Box office gross', breaks =
                         labels = c('2 million', '4 millio
                         low = 'blue', high = 'red')
```



# Putting it together



# Details

- Theme
- http://docs.ggplot2.org/current/theme.html

```
ggplot(data=imdb, aes(x=budget, y=imdb_score)) +
  geom_point(aes(colour=gross), alpha=0.5) +
  scale_x_continuous(breaks=c(1e6, 1e8),
                     labels=c('Low', 'Blockbuster'),
                     trans='log10') +
  geom_smooth() +
  scale_color_continuous(name='Box office gross', breaks =
                         labels = c('2 million', '4 millio
                         low = 'blue', high = 'red') +
  annotate('point', x=2e+08, y=8.3, colour='#D55E00', shape
  annotate('text', x=2e+08, y=8.7, label='Toy Story 3', fo
  annotate('point', x=210000000, y=5.7, colour='#009E73', s
  annotate('text', x=210000000, y=5.3, label='Transformers
  theme_bw() +
  labs(title='IMDB Movies', x='Budget', y='Rating') +
```

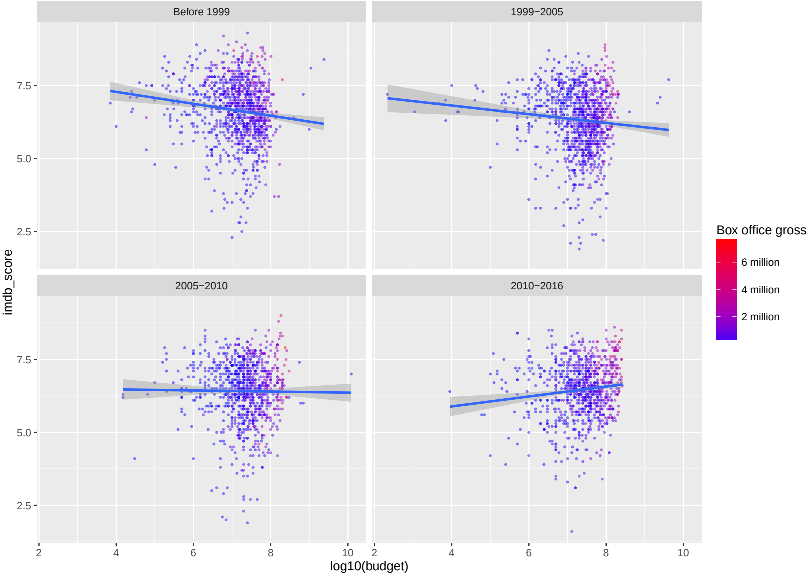
# Facet

- Similar plots for different periods with facet

```r
imdb$period <- cut(imdb$title_year, breaks=quantile(imdb$ti
                   labels=c('Before 1999', '1999-2005', '20
                   include.lowest=TRUE)
```

```r
ggplot(data=imdb, aes(x=log10(budget), y=imdb_score)) +
  geom_point(aes(colour=gross), alpha=0.5, size=0.5) +
  geom_smooth(method='lm') +
  scale_color_continuous(name='Box office gross', breaks =
                         labels = c('2 million', '4 millio
                         low = 'blue', high = 'red') +
  facet_wrap(~period)
```

# Facet


Before 1999
1999–2005
2005–2010
2010–2016
imdb_score
7.5
5.0
2.5
log10(budget)
2
4
6
8
10
Box office gross
6 million
4 million
2 million

# Multiple plots

- Multiplot function

http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)/

- gridExtra package

https://cran.r-project.org/web/packages/gridExtra/vignettes/arrangeGrob.html

# Resources

- Documentation

http://docs.ggplot2.org/current/

- Cheatsheet

https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf

- R Cookbook

http://www.cookbook-r.com/Graphs/

- Stackoverflow

http://stackoverflow.com/questions/tagged/ggplot2