title: ‘ggplot2 application (car:: Salaries)’
author: “cafepeng”
date: “2018年3月18日”
output: html_document
helpful link: [https://support.zendesk.com/hc/en-us/articles/203691016-Formatting-text-with-Markdown#topic_xqx_mvc_43__row_tf4_bmn_1n (https://support.zendesk.com/hc/en-us/articles/203691016-Formatting-text-with-Markdown#topic_xqx_mvc_43__row_tf4_bmn_1n)]
Dataset:

> The 2008-09 nine-month academic salary for Assistant Professors, Associate Professors and Professors in a college in the U.S. The data were collected as part of the on-going effort of the college’s administration to monitor salary differences between male and female faculty members.

- Variables:
  - rank: AssocProf, AsstProf, Prof
  - discipline:a factor with levels A (`theoretical’)or B (‘applied’ departments)
  - yrs.since.phd: years since PhD
  - yrs.service: years of service
  - sex: a factor with levels Female Male
  - salary: nine-month salary, in dollars

1. pull the dataset “Salaries” from ‘car’ package, and check the structure and so on.

```
library(car)
data01<-data.frame(Salaries)
str(data01)
```

```
## 'data.frame':    397 obs. of  6 variables:
##  $ rank         : Factor w/ 3 levels "AsstProf","AssocProf",..: 3 3 1 3 3 2 3 3 3 3 ...
##  $ discipline   : Factor w/ 2 levels "A","B": 2 2 2 2 2 2 2 2 2 2 ...
##  $ yrs.since.phd: int  19 20 4 45 40 6 30 45 21 18 ...
##  $ yrs.service  : int  18 16 3 39 41 6 23 45 20 18 ...
##  $ sex          : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 1 ...
##  $ salary       : int  139750 173200 79750 115000 141500 97000 175000 147765 119250 129000 ...
```

```
dim(data01)
```

```
## [1] 397   6
```

```
summary(data01)
```

```
##         rank     discipline yrs.since.phd    yrs.service        sex
##  AsstProf : 67   A:181      Min.   : 1.00   Min.   : 0.00   Female: 39
##  AssocProf: 64   B:216      1st Qu.:12.00   1st Qu.: 7.00   Male  :358
##  Prof     :266              Median :21.00   Median :16.00
##                             Mean   :22.31   Mean   :17.61
##                             3rd Qu.:32.00   3rd Qu.:27.00
##                             Max.   :56.00   Max.   :60.00
##      salary
##  Min.   : 57800
##  1st Qu.: 91000
##  Median :107300
##  Mean   :113706
##  3rd Qu.:134185
##  Max.   :231545
```

```
#especially check the variables I am interested in
summary(data01$rank)
```

```
##  AsstProf AssocProf      Prof
##        67        64       266
```

```
summary(data01$salary)
```

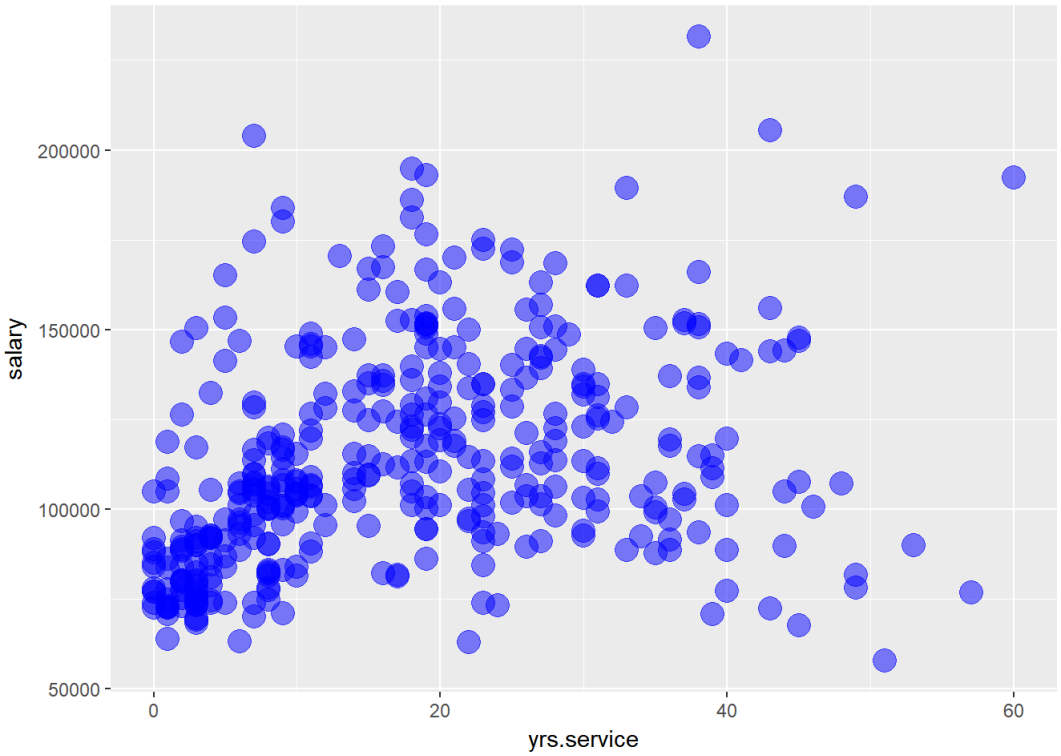
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   57800   91000  107300  113706  134185  231545
```

```
summary(data01$sex)
```

```
## Female   Male
##     39    358
```

- **Q1. If income increases as the year of service increases?**

```
#Load the packages needed for data visualization (note: eval=FALSE prevents code from being evaluated)
library(ggplot2)
#note: alpha refers to the degree of transparancy so that the points would not block other points
pic1<-ggplot(data01)+geom_point(mapping=aes(x=yrs.service,y=salary),alpha=1/2,color="blue",size=5)
pic1
```



Q1 Finding: as years of service increases, the salaries increase as well

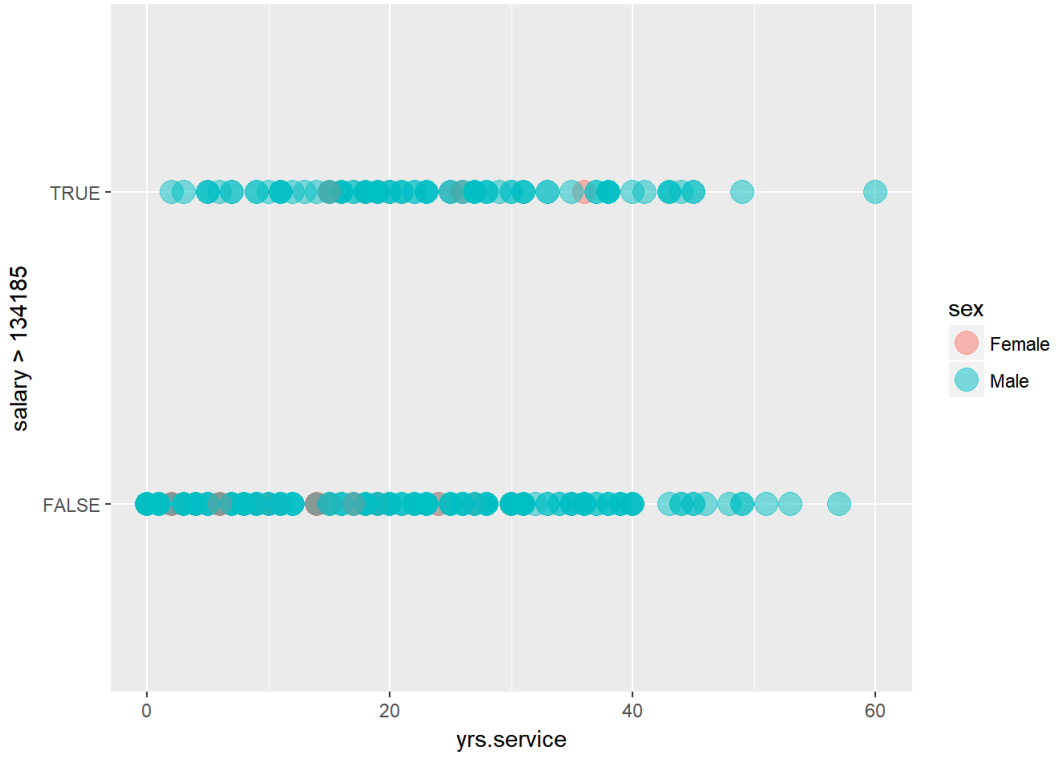
- **Q2.is there a difference in salaries between males and females overall?**

```
# add another aesthetic measure "sex"
library(ggplot2)
pic2<-ggplot(data01)+geom_point(mapping=aes(x=yrs.service,y=salary,color=sex),alpha=1/2,size=5)
pic2
```

Q2 Findings: it seems that the salary range is relatively limited to females

- **Q3_1. following Q2, is it possible that the difference btw males and females gets larger in higher income range(e.g. higher than Q3)?**

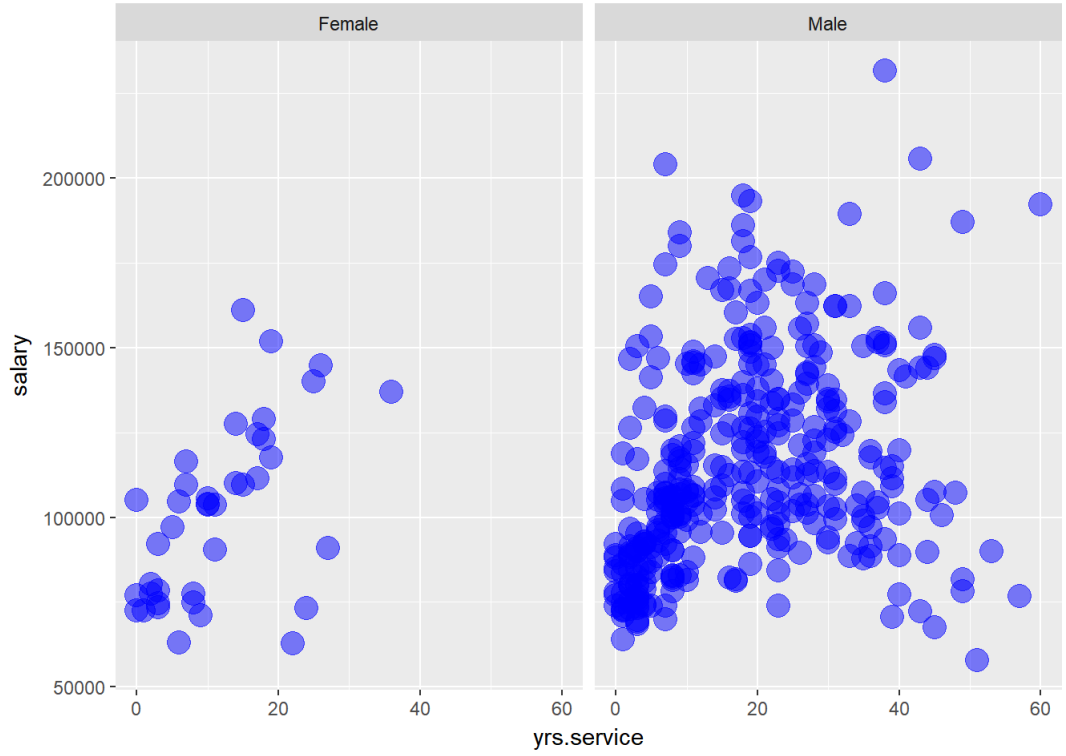
```
library(ggplot2)
pic3_1<-ggplot(data01)+geom_point(mapping=aes(x=yrs.service,y=salary>134185,color=sex),alpha=1/2,size=5)
pic3_1
```



Q3-1 Finding: due to the relative small sample of female, this picture is relatively meaningless

- **Q3_2.Use “facet”" setting to draw two diagrams that represents females and males seperately**

```
library(ggplot2)
pic3_2<-ggplot(data01)+geom_point(mapping=aes(x=yrs.service,y=salary),color="blue",alpha=1/2,size=5)+facet_wrap(~sex)
pic3_2
```

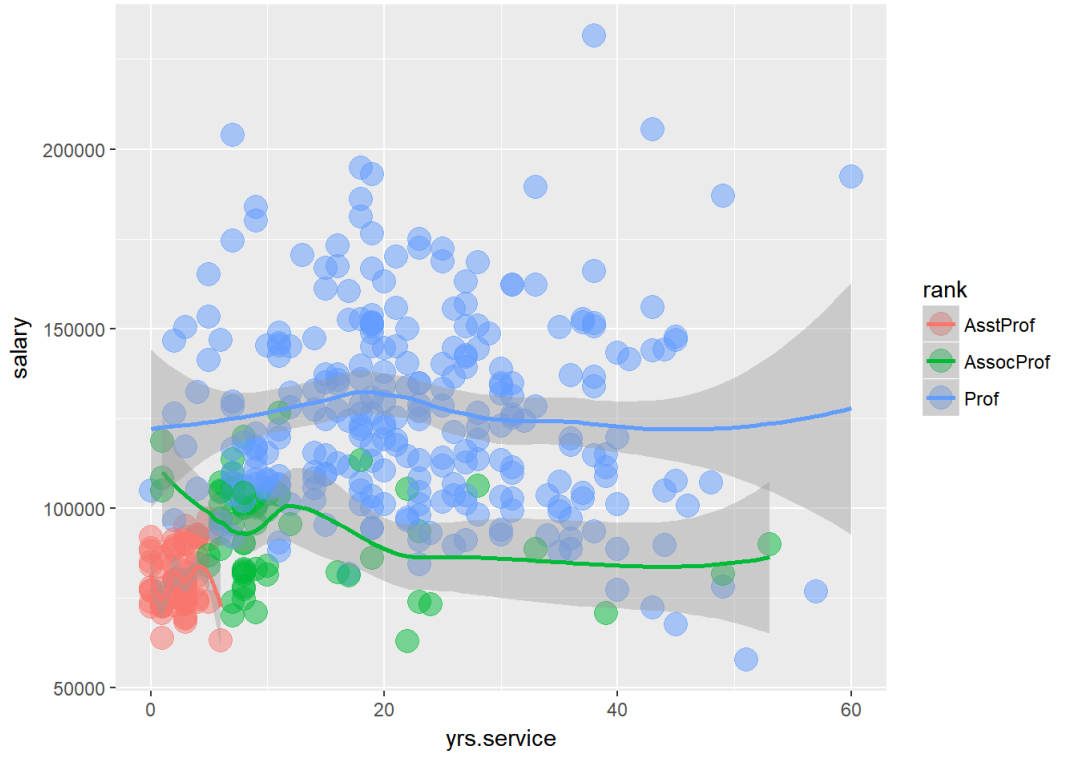


Q3_2 Finding: females' salary range tends to be relatively limited compared to males

- **4_1.Does the salary differs in terms of ranks?**

```
library(ggplot2)
pic4_1<-ggplot(data01,mapping=aes(x=yrs.service,y=salary))+
  geom_point(
      mapping=aes(color=rank),alpha=1/2,size=5)+
  geom_smooth(
      mapping=aes(color=rank))
pic4_1
```

```
## `geom_smooth()` using method = 'loess'
```



Q4_1 Finding: higher the rank, higher the variance in salaries

Note: we can use the command 'ggsave' to save the diagram. `ggsave(filename="pic4_1.pdf",plot=pic4_1)`

- **Q4_2: Take a closer look at the highest rank "professor" by plotting its line only**

```
library(ggplot2)
require(dplyr) # filter is part of dplyr package
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

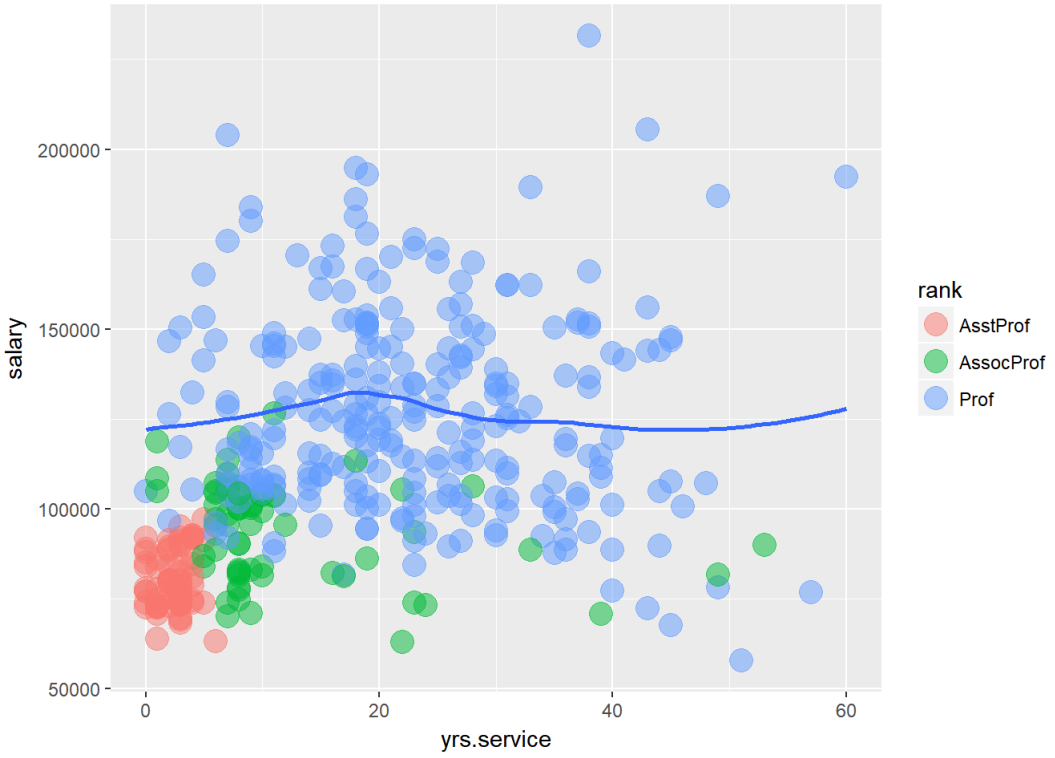
```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
pic4_2<-ggplot(data01,mapping=aes(x=yrs.service,y=salary))+geom_point(mapping=aes(color=rank),alpha=1/2,size=5)+geom_s
mooth(data=filter(data01,rank=="Prof"), se=FALSE)
pic4_2
```

```
## `geom_smooth()` using method = 'loess'
```



Q4_2 Finding: higher the rank, highe the range in salaries