Lab1

Melwin Poovakottu

Friday, January 20, 2017

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
getwd()
```

```
## [1] "C:/Users/Melwin/Desktop/Data Science files/UC Berkeley/W203 Stats/W203_Assignments/Lab_1"
```

```
library(car)
library(ggplot2)
library(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
```

```
load("ceo_w203.RData")
summary(CEO)
```

```
##      salary            age           college            grad
##  Min.   : 100.0   Min.   :21.00   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 467.0   1st Qu.:51.00   1st Qu.:1.0000   1st Qu.:0.0000
##  Median : 697.0   Median :57.00   Median :1.0000   Median :1.0000
##  Mean   : 852.9   Mean   :55.78   Mean   :0.9622   Mean   :0.5514
##  3rd Qu.:1101.0   3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :5299.0   Max.   :86.00   Max.   :1.0000   Max.   :1.0000
##      comten          ceoten           profits          mktval
##  Min.   : 2.00   Min.   : 0.000   Min.   :-463.0   Min.   :   -1
##  1st Qu.: 9.00   1st Qu.: 3.000   1st Qu.:  33.0   1st Qu.:  567
##  Median :21.00   Median : 5.000   Median :  57.0   Median : 1200
##  Mean   :21.66   Mean   : 7.681   Mean   : 199.2   Mean   : 3450
##  3rd Qu.:33.00   3rd Qu.:11.000   3rd Qu.: 195.0   3rd Qu.: 3200
##  Max.   :58.00   Max.   :37.000   Max.   :2700.0   Max.   :45400
```

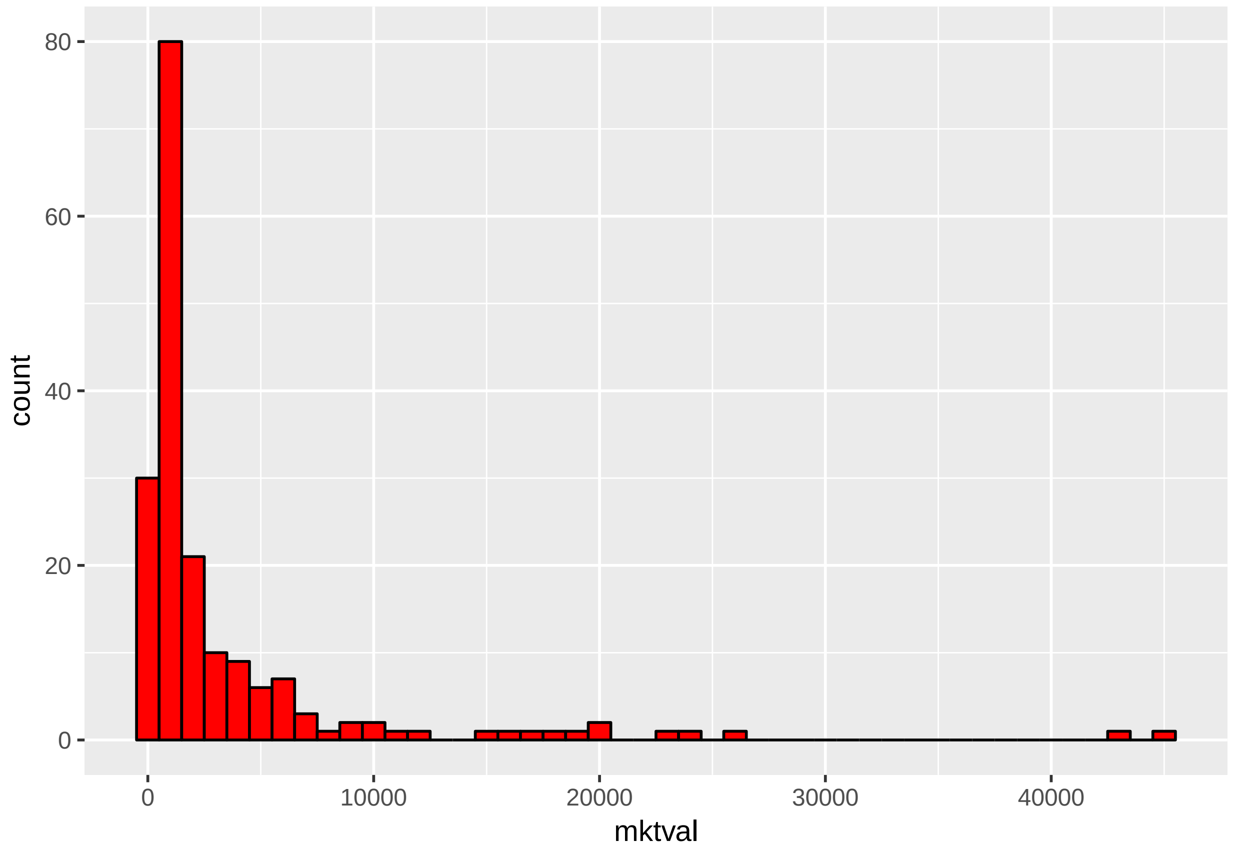
The market value has continous variable. While running a summary command you see there is big difference in the mean(=3450) and Median(=1200). This means that the data is positively skewed to the right with a lot data points close to the median and a relatively high number of of outliers with higher market value.

When we see the boxplot it is clear that there are a many values which are above 1.5 times the 3 quatrile, causing the positive skew.

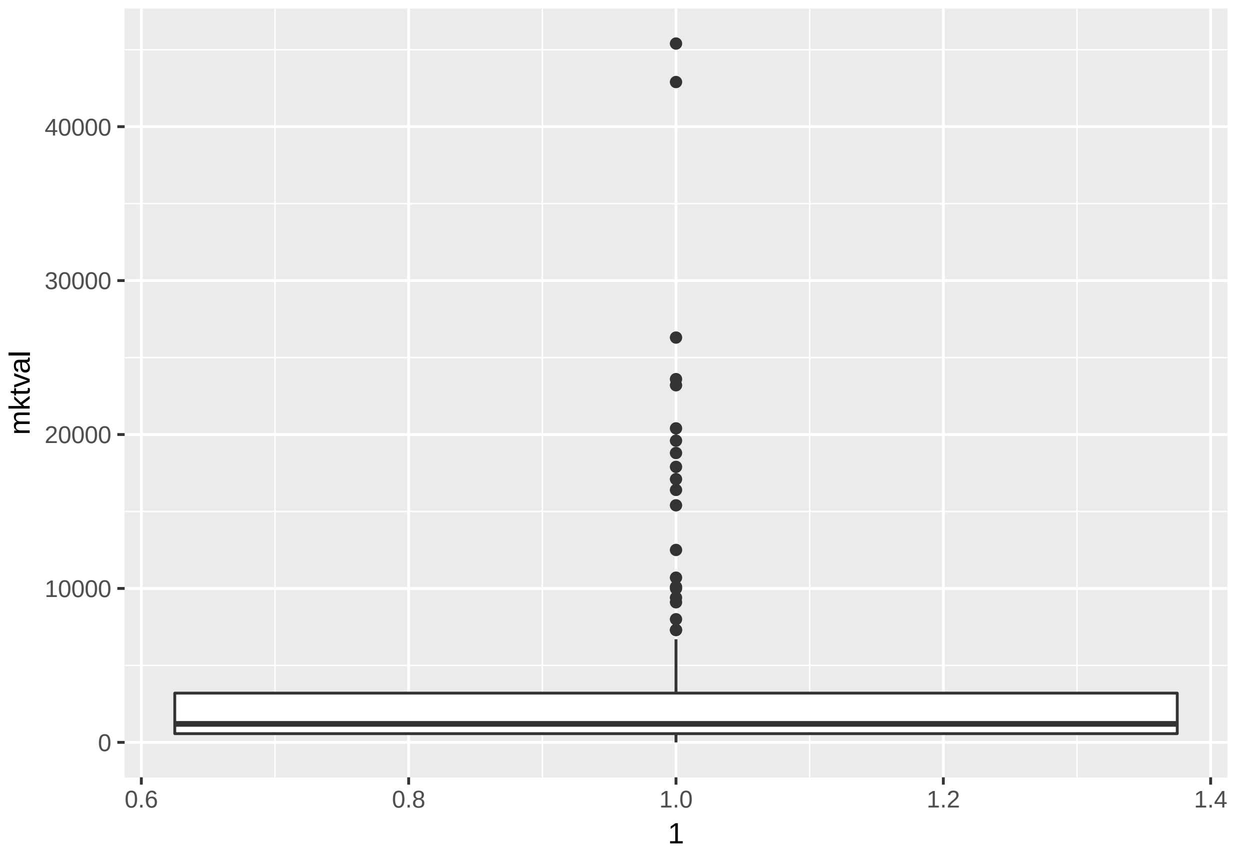
```
summary(CEO$mktval)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##      -1     567    1200    3450    3200   45400
```

```
ggplot(CEO,aes(mktval,color = as.factor(grad)))+ geom_histogram(binwidth = 1000,fill ="red",col="black"
```



```
ggplot(CEO,aes(y=mktval,x=1))+ geom_boxplot()
```



We see that there are some negative market value data points. These data points also have the profits as -1. This could potentially instances of missing data points and default value of -1 is recorded. So we will be considering them as na.

```
CEO[CEO$mktval==-1,]
```

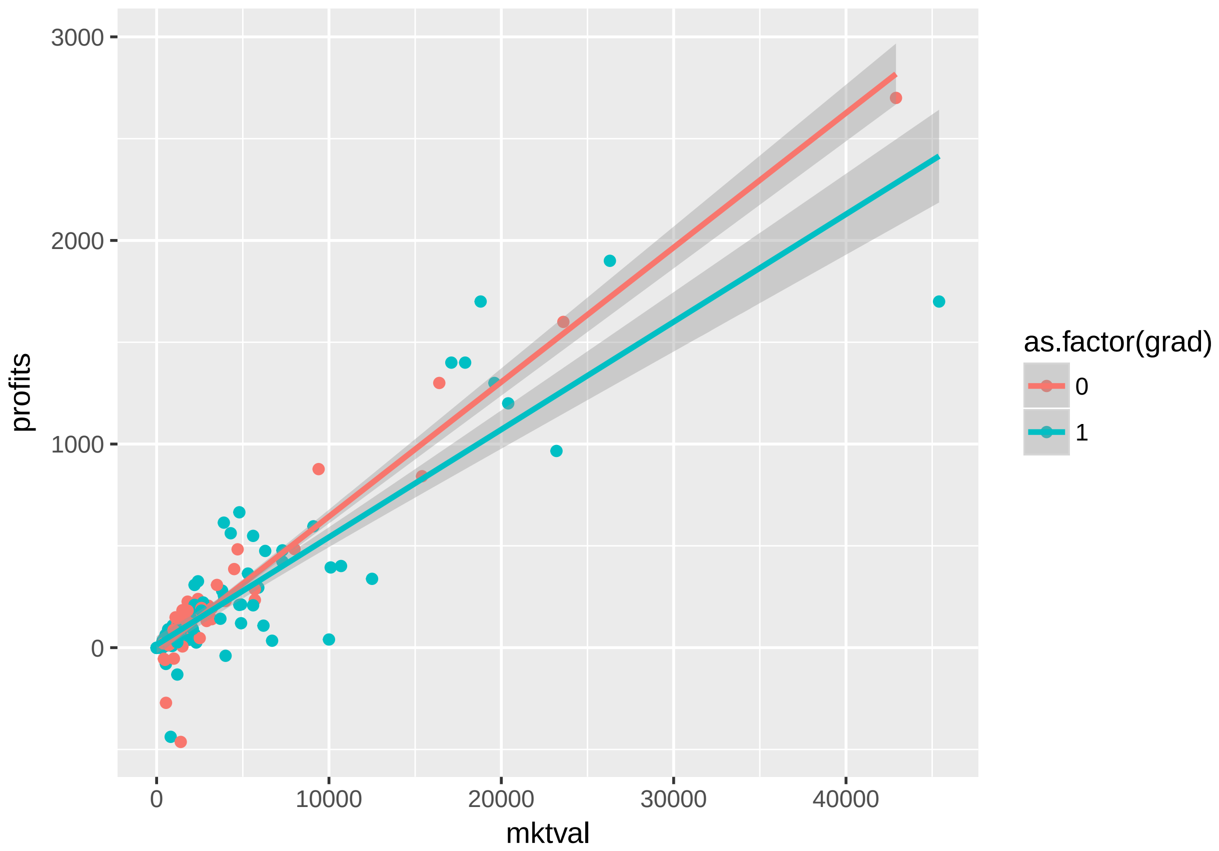
```
##     salary age college grad comten ceoten profits mktval
## 178    379  55       1    1      4      2      -1     -1
## 179    677  31       1    1      3      1      -1     -1
## 182    637  45       1    1      3      1      -1     -1
## 181    873  61       1    1      3      1      -1     -1
## 180    173  55       1    1      3      1      -1     -1
```

We can see a high correlation between market value and profits, which is expected for most companies. Similary we see a good correlation between market value and salary. Also we see that the if the CEO has a graduate degree he is more likely to get a better salary than without a graduate degree On ploting the salary and profits against mktval, we see there are few data point that are outliers which are affecting the regression line.

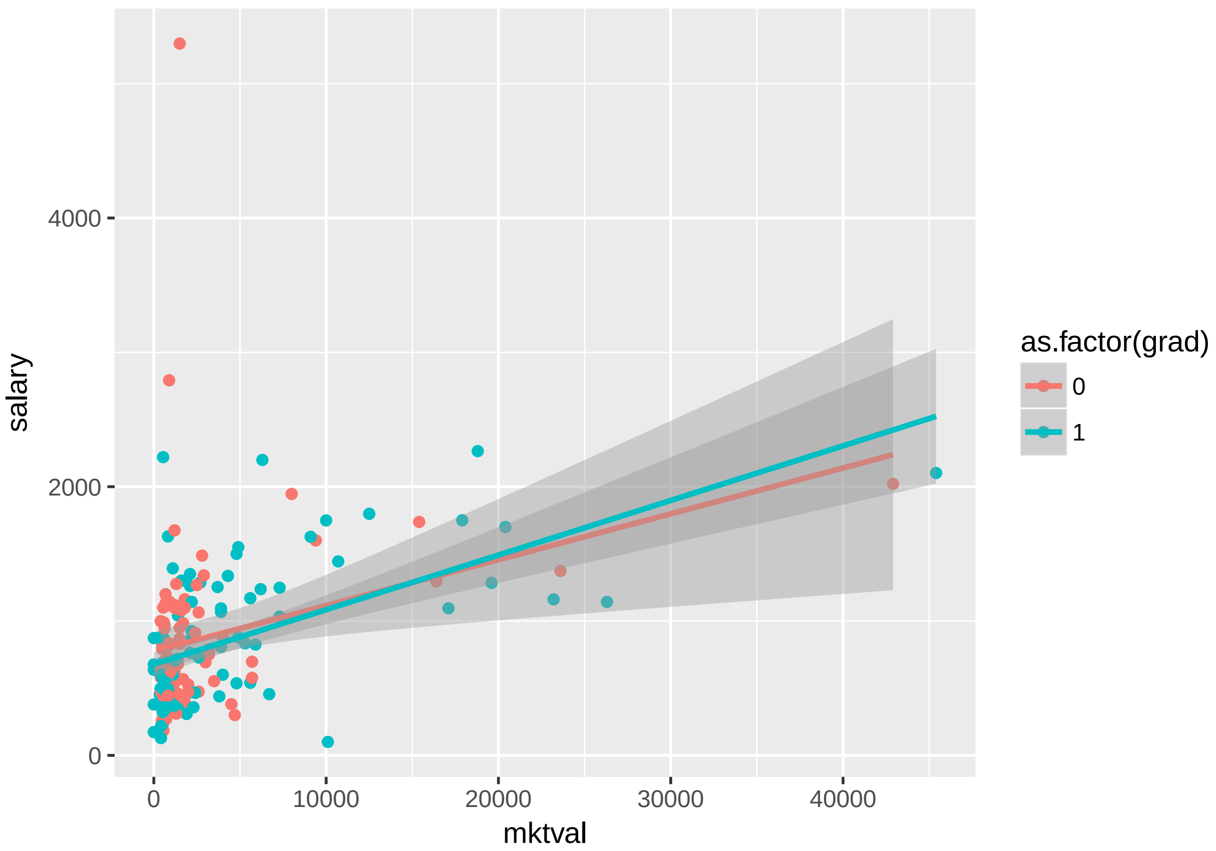
```
cor(CEO$mktval,CEO, use="complete.obs")
```

```
##         salary       age     college       grad    comten     ceoten
## [1,] 0.4119486 0.1308995 0.001185881 0.09848477 0.1633913 0.02681392
##        profits mktval
## [1,] 0.9190233      1
```

```
ggplot(CEO, aes(x = mktval, y = profits, color = as.factor(grad))) +geom_point() + stat_smooth(method =
```



```
ggplot(CEO, aes(x = mktval, y = salary, color = as.factor(grad))) +geom_point() + stat_smooth(method =
```



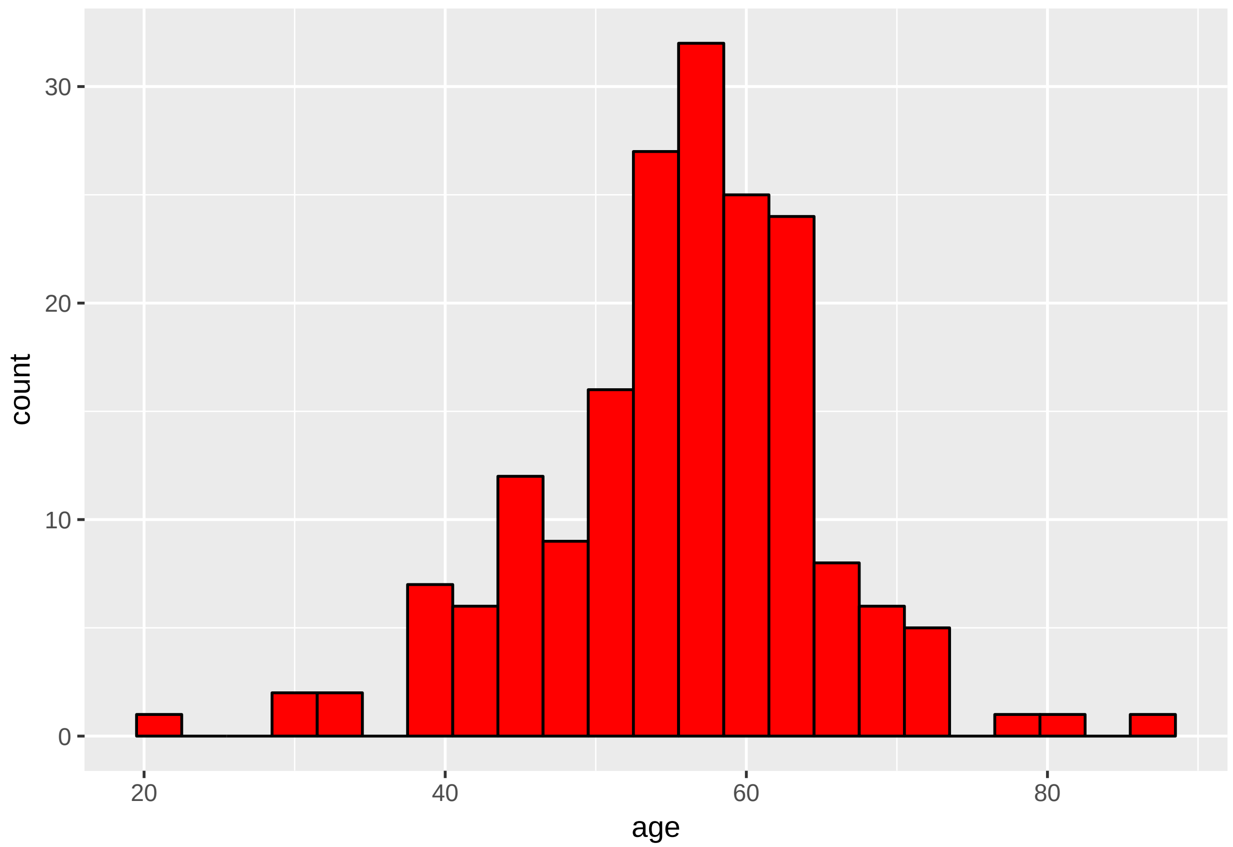
Age

Age is a discrete variable. Looking at the histogram plot, we can see that the variable is close to a normally distributed curve with the mean(55) and median(57). But you can see a drastic drop after the age of 65. This likey co-responds to the age of retirement in the US.

```
summary(CEO$age)
```

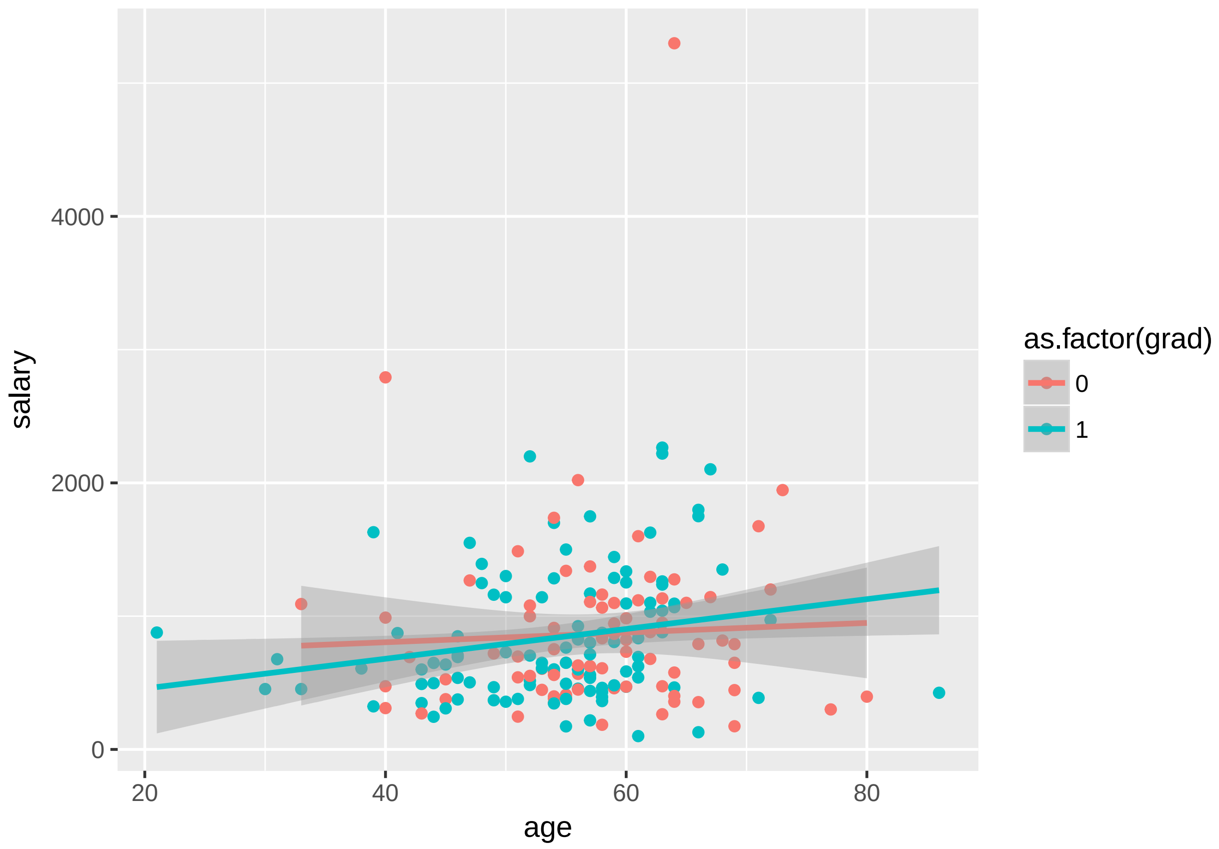
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   21.00   51.00   57.00   55.78   61.00   86.00
```

```
ggplot(CEO,aes(age,color = as.factor(grad))+ geom_histogram(binwidth = 3,fill ="red",col="black")
```



If we simply plot a graph of age vs salary do not get any particular co-relation. The corelatoin is also very low(0.13) We do see that if you have a grad degree you are likely to get paid more as you in advance your career.

```
ggplot(CEO,aes(x=age,y=salary,color = as.factor(grad))+ geom_point()+ stat_smooth(method = "lm")
```

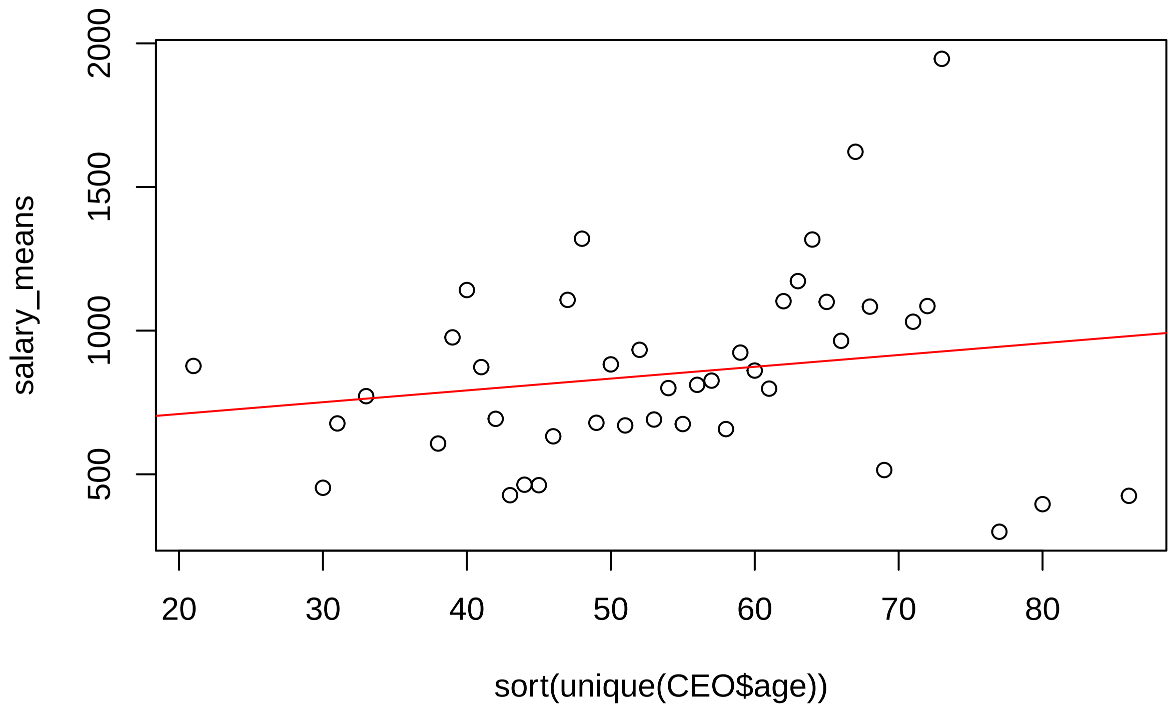


```
cor(CEO$age,CEO$salary)
```

```
## [1] 0.130081
```

Let us examine age variable closely. Since age is a discreet variable let us group the salary's of all the CEOs of the same age. We can take mean of all the salary's in each group. On ploting the mean salary against the age we see an interesting trend We see that within the age range 29 to 75 there might be a corelation between average salary and age. This is also the range which has the maxium data points

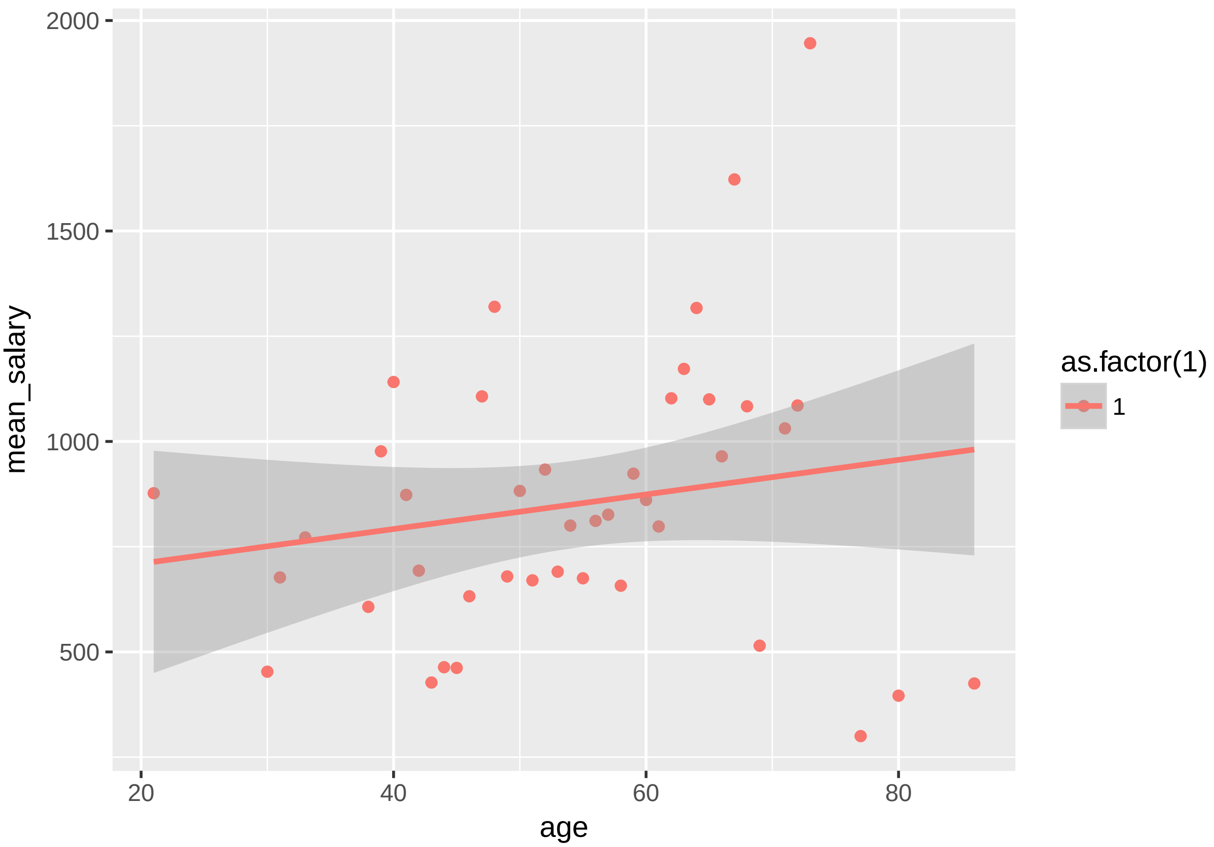
```
salary_means= by(CEO$salary,CEO$age,mean)
plot(sort(unique(CEO$age)), salary_means)
abline(lm(salary_means~sort(unique(CEO$age))),col="red")
```



```
cor(salary_means,sort(unique(CEO$age)))
```

```
## [1] 0.17776
```

```
grouped = group_by(CEO, age)
mean_salary_age = summarise(grouped, mean_salary = mean(salary, na.rm = T))
ggplot(mean_salary_age,aes(x=age,y=mean_salary,color = as.factor(1)))+ geom_point()  + stat_smooth(meth
```

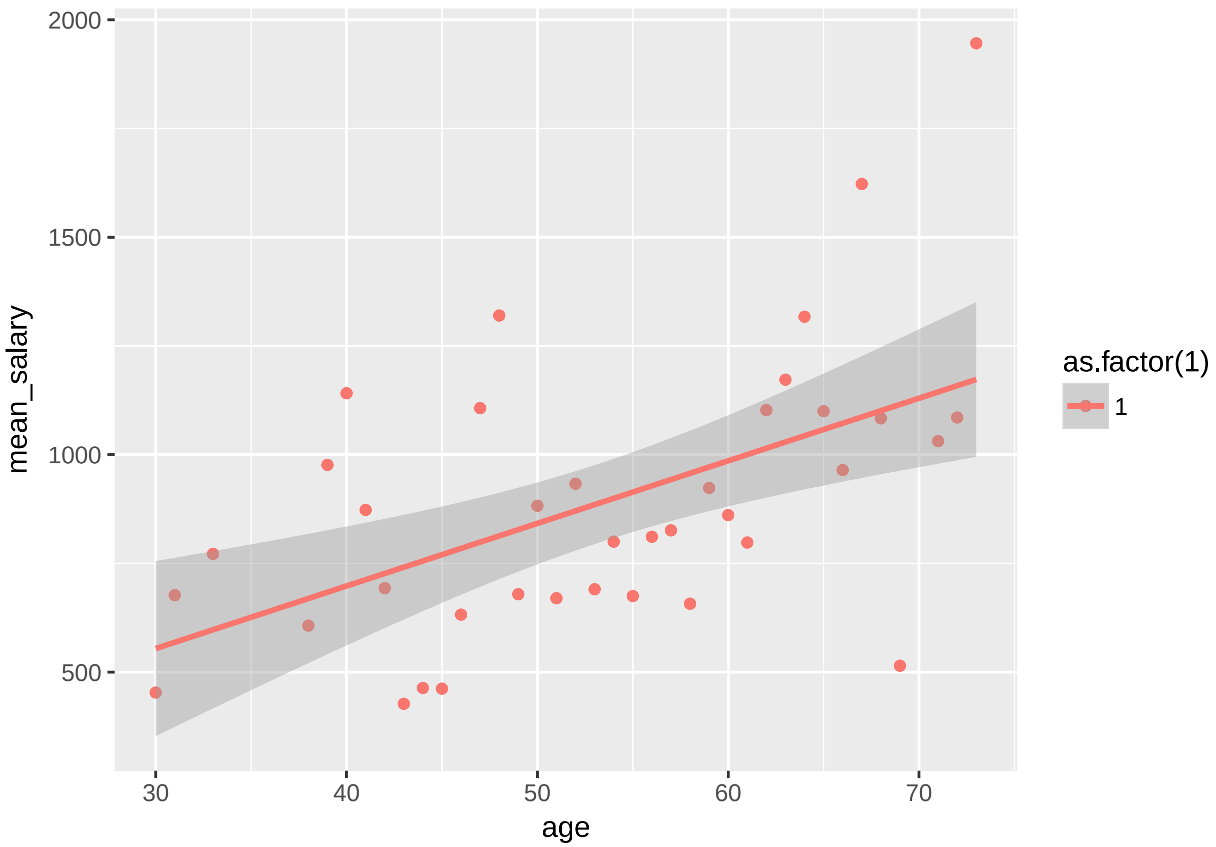


```
cor(mean_salary_age)
```

```
##                  age mean_salary
## age          1.00000     0.17776
## mean_salary 0.17776     1.00000
```

Let Narrowing age from 29 to 75 We see a high co-relation between mean salary and the age within the age range 29 to 75

```
narrowed_mean_salary_age=mean_salary_age[mean_salary_age$age<=75 & mean_salary_age$age>=29,]
ggplot(narrowed_mean_salary_age,aes(x=age,y=mean_salary,color = as.factor(1)))+ geom_point()  + stat_smo
```



```
cor(narrowed_mean_salary_age$age,narrowed_mean_salary_age$mean_salary)
```

```
## [1] 0.5321218
```