kuis2_G64190069_Rizal_Mujahiddan

Rizal Mujahiddan

3/12/2022

Soal Nomor 1

```
library(MASS)
library(ggplot2)
library(visdat)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
```

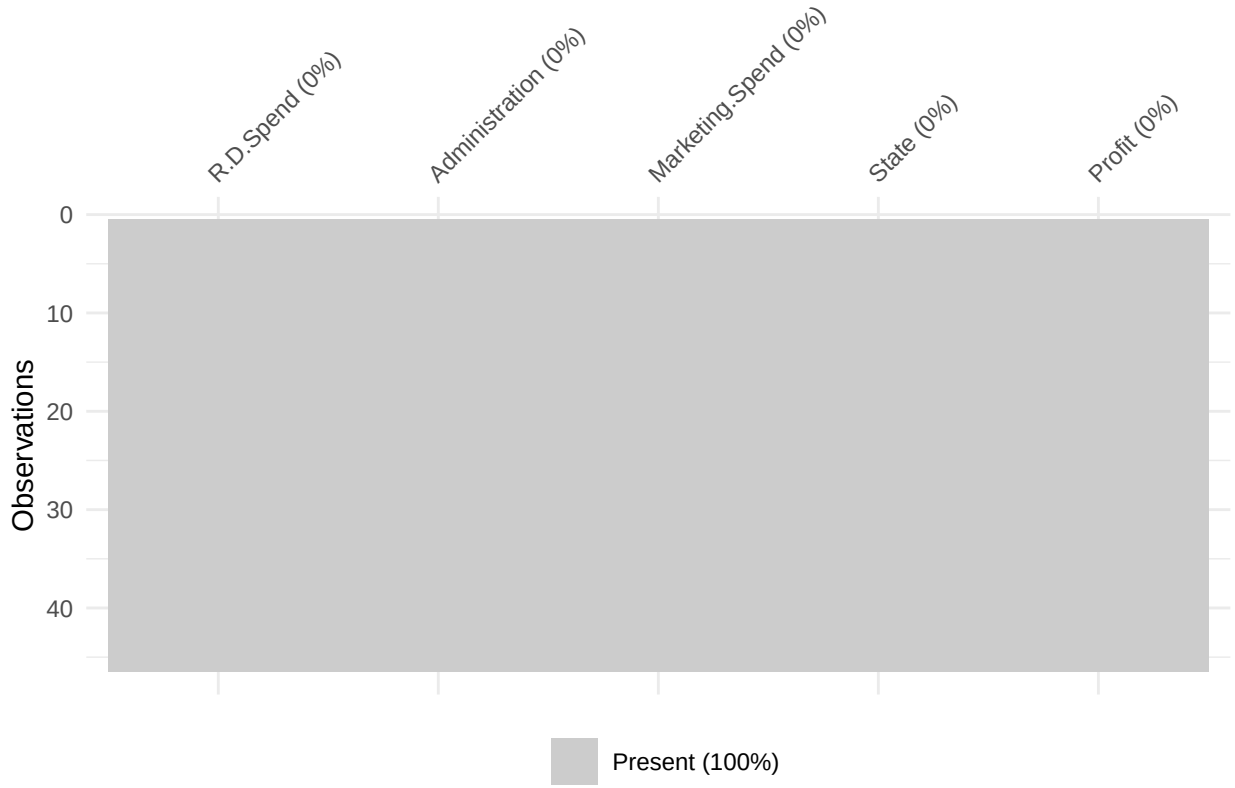
```
library(rcompanion)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
dataku <- read.csv("Kuis UTS 2.csv",sep = ";")
head(dataku)
```

```
##   R.D.Spend Administration Marketing.Spend      State      Profit
## 1  165349,2       136897,8        471784,1   New York 369646,1127
## 2  162597,7      151377,59       443898,53 California 367841,9428
## 3 153441,51      101145,55       407934,54    Florida 365002,5152
## 4 144372,41      118671,85       383199,62   New York 334531,3795
## 5 142107,34       91391,77       366168,42    Florida  276184,314
## 6  131876,9       99814,71       362861,36   New York 246462,1176
```

```
vis_miss(dataku)
```



bisa dilihat bhawa datanya tidak memiliki missing value, maka bisa langsung dieksplorasi , tapi sebelumnya kita memperbaiki tipe data tersebut ya

```
str(dataku)
```

```
## 'data.frame':    46 obs. of  5 variables:
##  $ R.D.Spend      : chr  "165349,2" "162597,7" "153441,51" "144372,41" ...
##  $ Administration : chr  "136897,8" "151377,59" "101145,55" "118671,85" ...
##  $ Marketing.Spend: chr  "471784,1" "443898,53" "407934,54" "383199,62" ...
##  $ State          : chr  "New York" "California" "Florida" "New York" ...
##  $ Profit         : chr  "369646,1127" "367841,9428" "365002,5152" "334531,3795" ...
```

```
# Diprogram ini, penulis mengubah tipe datanya ya
# mengubah koma menjadi titik agar menjadi double ya atau numeric

num_kol <- c("R.D.Spend","Administration","Marketing.Spend","Profit")
for(i in num_kol){
  dataku[[i]] <- sub(",",".",dataku[[i]])
  dataku[[i]] <- as.numeric(dataku[[i]])
}

head(dataku)
```

```
##   R.D.Spend Administration Marketing.Spend      State   Profit
```

```
## 1  165349.2      136897.80        471784.1   New York 369646.1
## 2  162597.7      151377.59        443898.5 California 367841.9
## 3  153441.5      101145.55        407934.5    Florida 365002.5
## 4  144372.4      118671.85        383199.6   New York 334531.4
## 5  142107.3       91391.77        366168.4    Florida 276184.3
## 6  131876.9       99814.71        362861.4   New York 246462.1
```

String Prerocessing

Setelah diubah pada bagian numericnya, maka alangkah lebih baiknya String bisa kita cek terlebih dahulu apakah bisa dilakukan faktor data gitu

```
# dikarenakan hanya satu kolom, maka kita ubah saja satu kolom tanpa buat vector
# kolom tersebut ya
#ini memastikan tidak ada typo pada kolom tersebut
print(unique(dataku$State))
```

```
## [1] "New York"   "California" "Florida"
```

Pemfaktoran

Ternyata tidak ada typo dalam data Tersebut ya langsung kita faktorkan

```
dataku$State <- factor(dataku$State)
head(dataku)
```

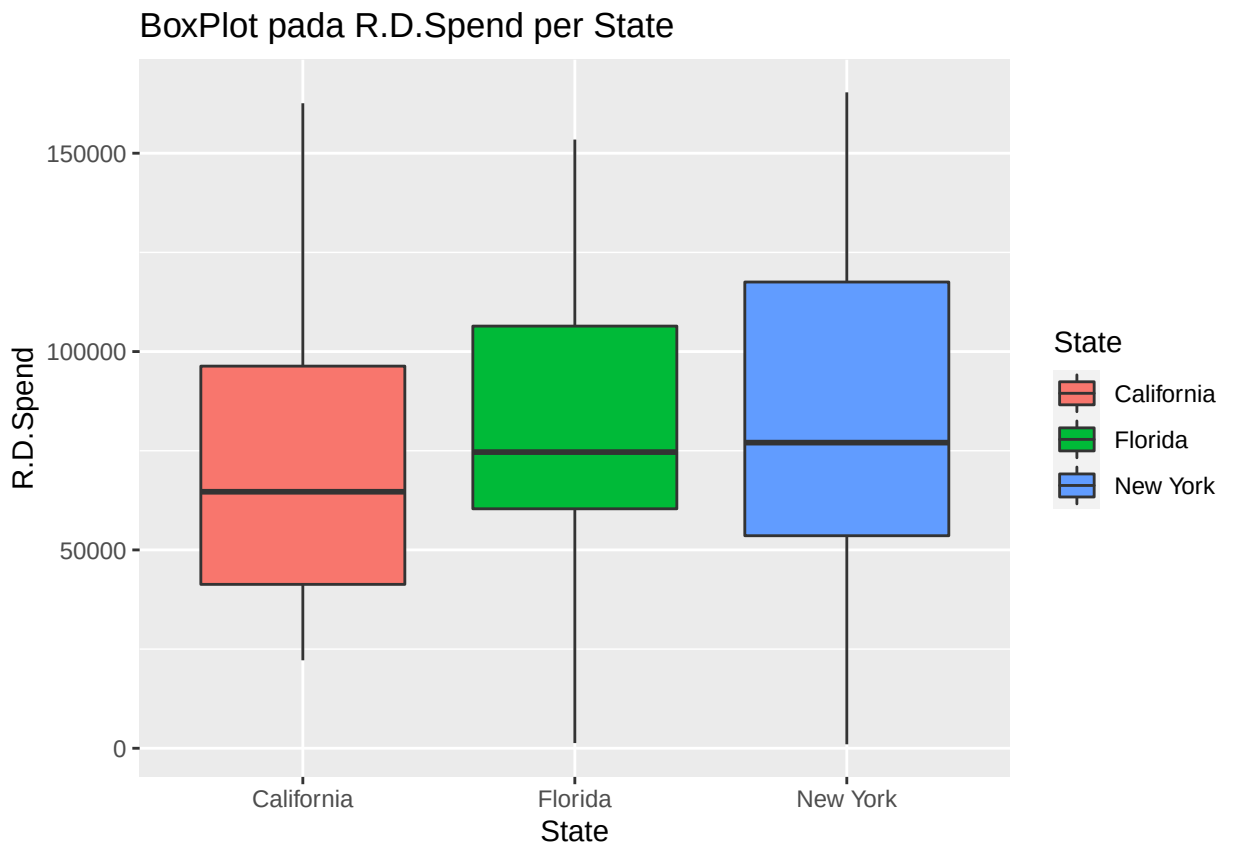
```
##   R.D.Spend Administration Marketing.Spend      State   Profit
## 1  165349.2      136897.80        471784.1   New York 369646.1
## 2  162597.7      151377.59        443898.5 California 367841.9
## 3  153441.5      101145.55        407934.5    Florida 365002.5
## 4  144372.4      118671.85        383199.6   New York 334531.4
## 5  142107.3       91391.77        366168.4    Florida 276184.3
## 6  131876.9       99814.71        362861.4   New York 246462.1
```

Eksplorasi data

R.D.Spend

Boxplot

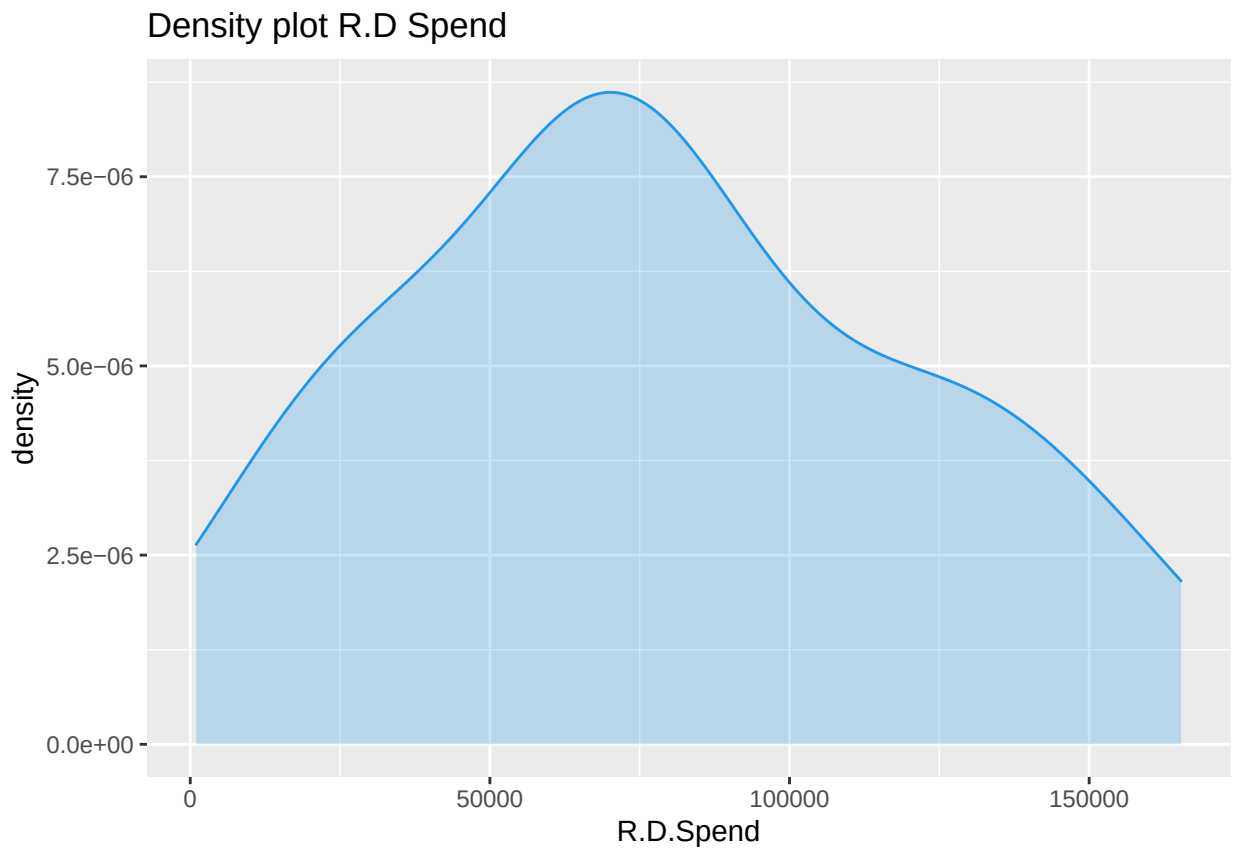
```
ggplot(dataku,aes(x=State,y=R.D.Spend,fill=State)) +
  ggtitle("BoxPlot pada R.D.Spend per State ")+
  geom_boxplot()
```



Jika kita lihat, Tidak ada outliernya, berarti kira kira Tidak ada biaya tambahan yang berlebih atau yang berkurang dari suatu wilayah tersebut ya didalam R & D Spend di setiap state, meskipun ada Perbedaan sedikit ya, Newyork lebih besar dikarenakan newyork sendiri menjadi ibaratnya kota ekonomi yang lebih pesat dan banyak kantor disana seperti Kota Jakarta kalau di Indonesia

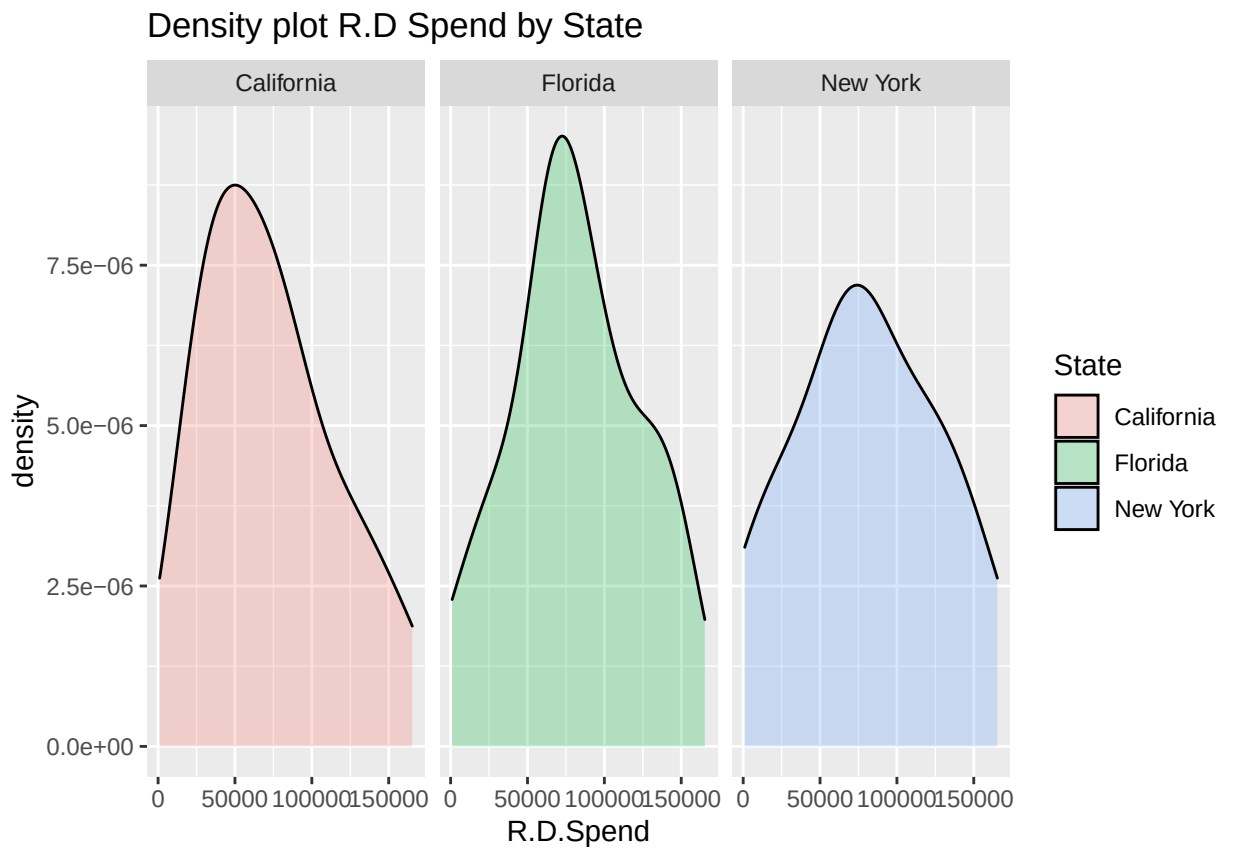
Density plot

```
ggplot(dataku,aes(x=R.D.Spend)) +
  geom_density(color=4,fill=4,alpha=0.25) +
  ggtitle("Density plot R.D Spend")
```



Sekilas, R.D Spend ini menyebar normal yang artinya, bahwa biaya tersebut bisa dikatakan wajar, dan tidak ada keanehan dalam R.D spend tersebut, maka ini jika ditinjau keseluruhan

```
ggplot(dataku,aes(x=R.D.Spend)) +
  geom_density(mapping=aes(fill=State),alpha=0.25) +
  ggtitle("Density plot R.D Spend by State") +
  facet_wrap(~ State, ncol = 3)
```



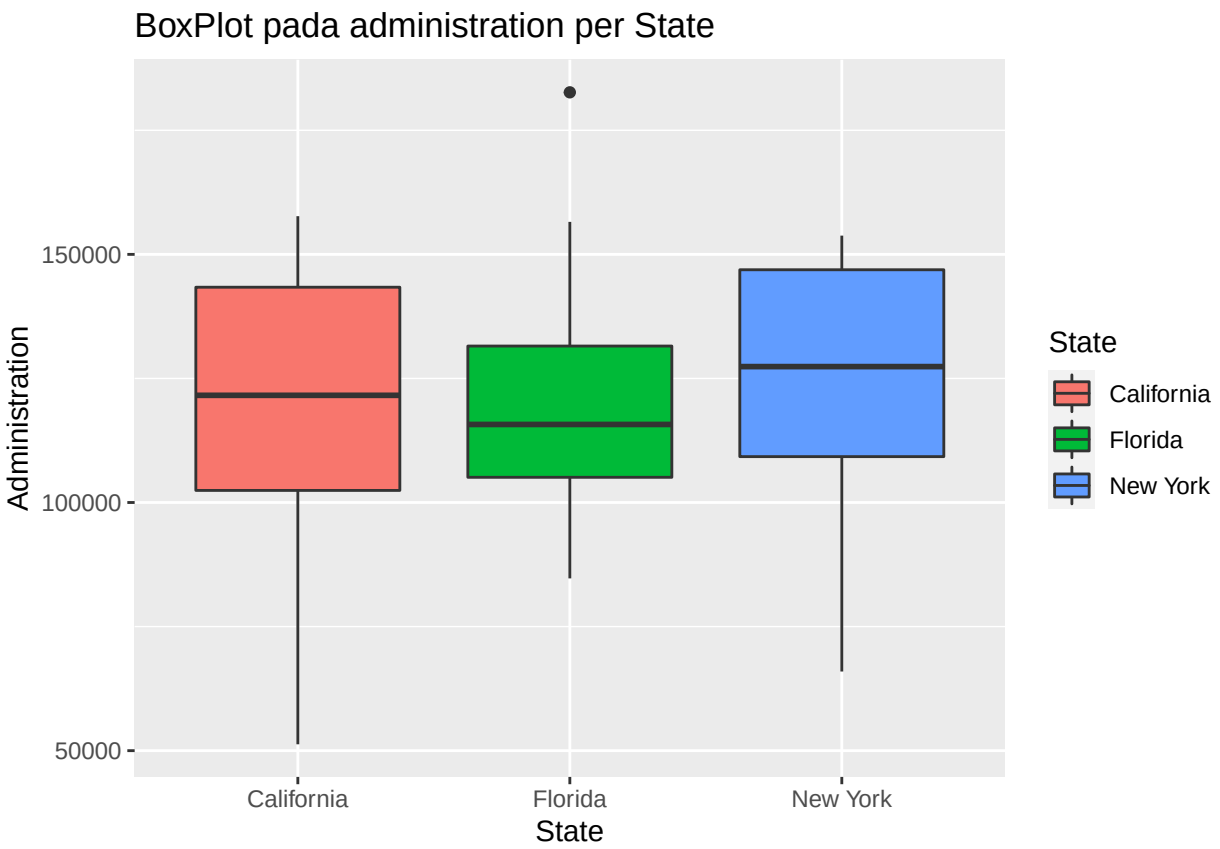
Disini bisa kita lihat cenderung normal, berarti tidak ada keanehan pada data tersebut , Kecuali pada California, tidak tepat ditengah (Subjektif)

```
data_california <- dataku %>% filter(State == "California")
data_new_york <- dataku %>% filter(State == "New York")
data_florida <- dataku %>% filter(State == "Florida")
```

Administration

Boxplot

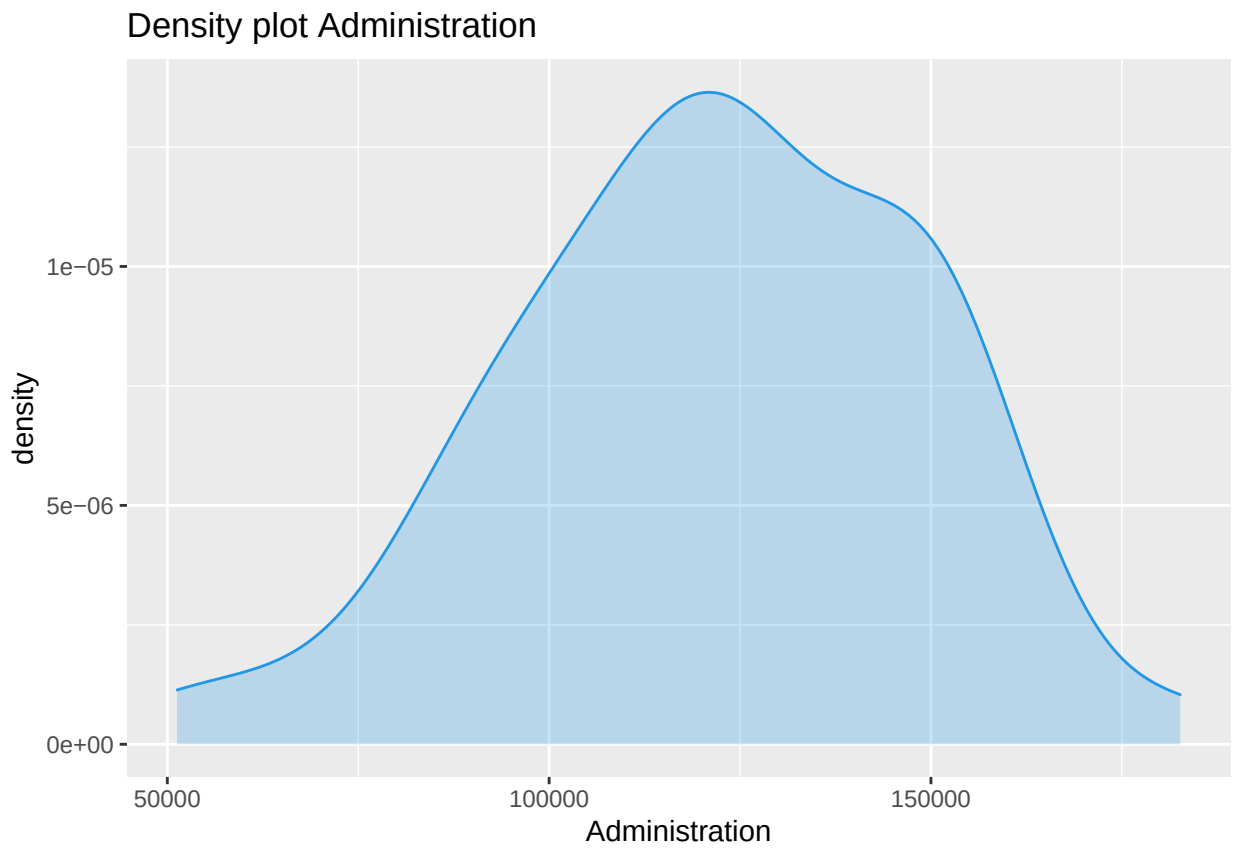
```
ggplot(dataku,aes(x=State,y=Administration,fill=State)) +
  ggtitle("BoxPlot pada administration per State ")+
  geom_boxplot()
```



Disini , ada biaya Administrasi , untuk rata rata hampir mendekati sama untuk masing masing state. Tetapi masalahnya ada di florida. Kemungkinan ada praktik Fraud dalam pembiayaan, atau jika kita berprasanka baik, ada kesalahan dalam masuk data

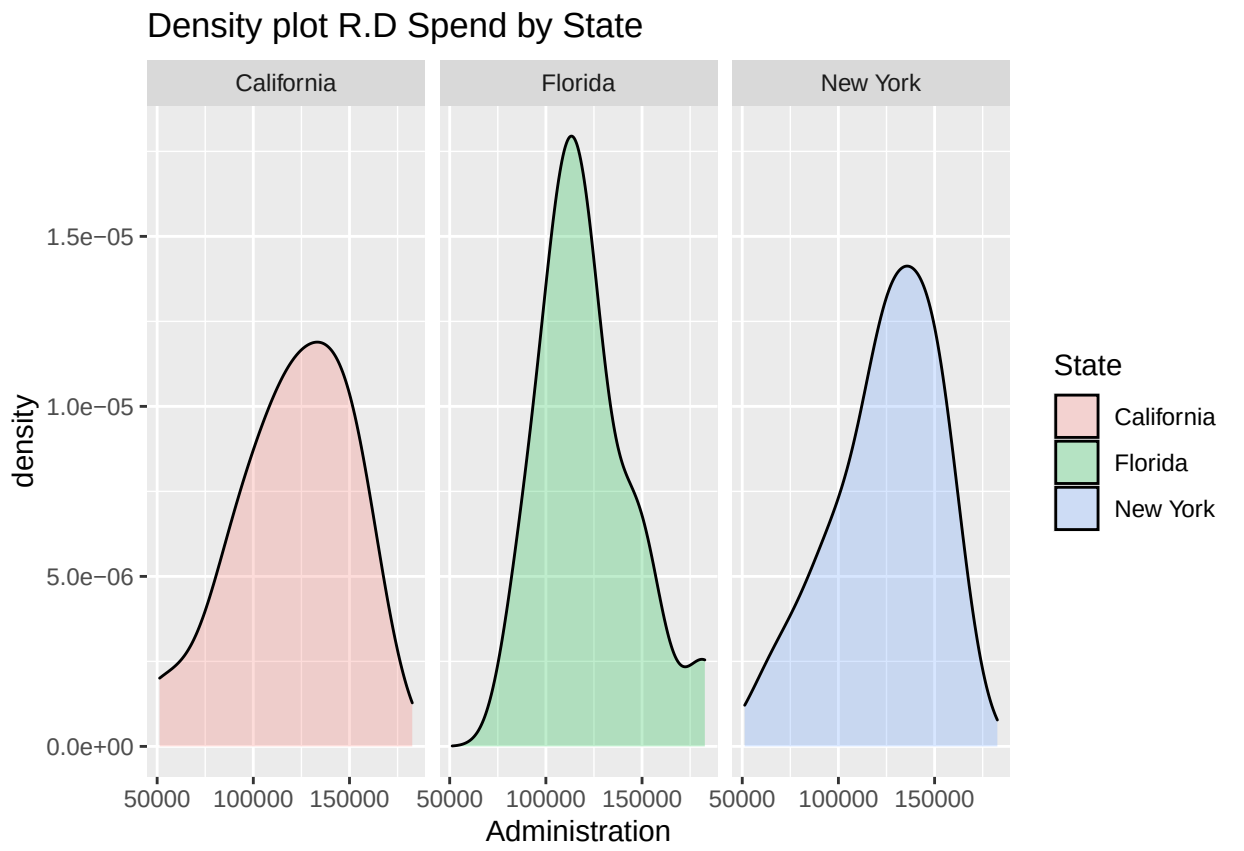
Density plot

```
ggplot(dataku,aes(x=Administration)) +
  geom_density(color=4,fill=4,alpha=0.25) +
  ggtitle("Density plot Administration")
```



di Administration ini, persebarannya normal tetapi di gelembung di sebelah kanan dan mengembung, maka bisa dipastikan bahwa administrasi mungkin tidak begitu rata dilapangan, berarti ada suatu halmungkin dari biaya berbeda di suatu daerah, maka kita cek lagi

```
ggplot(dataku,aes(x=Administration)) +
  geom_density(mapping=aes(fill=State),alpha=0.25) +
  ggtitle("Density plot R.D Spend by State") +
  facet_wrap(~ State, ncol = 3)
```



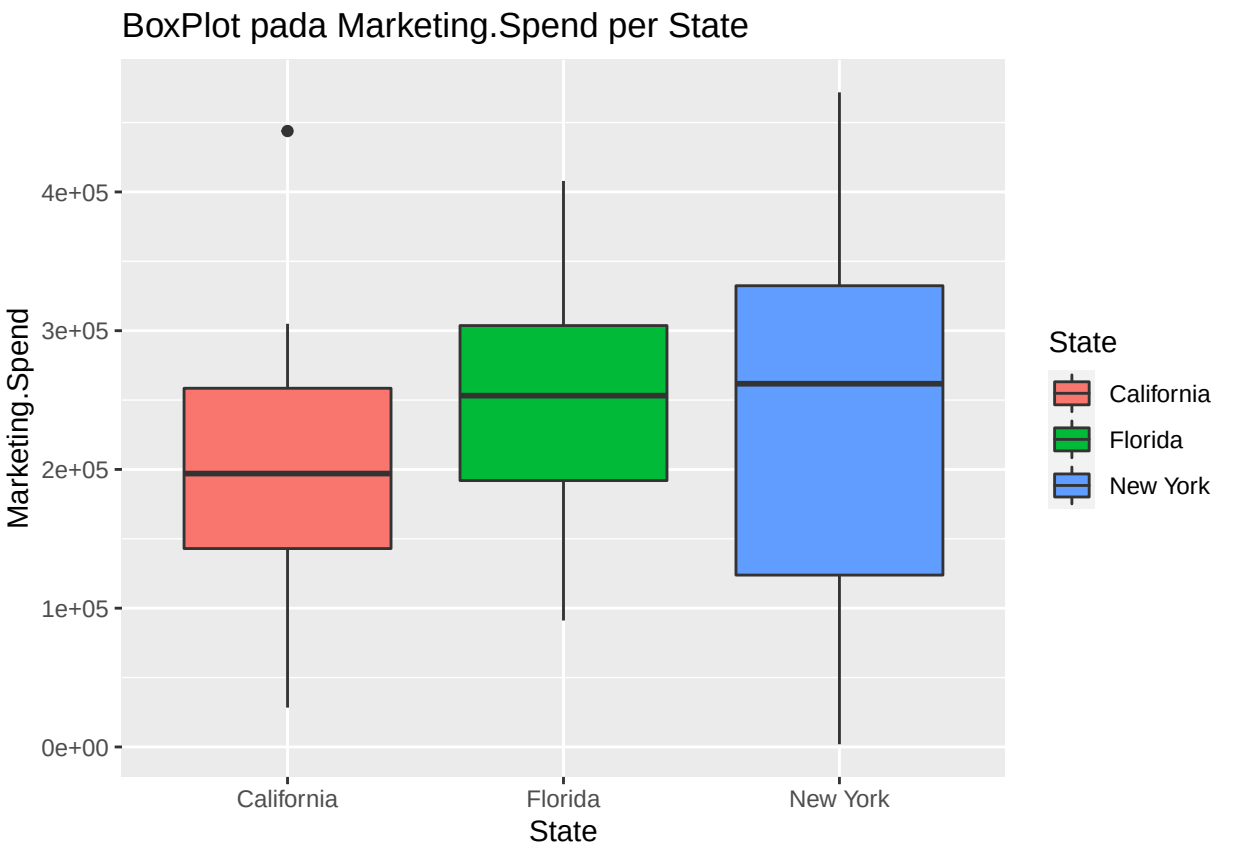
UNtuk florida, sangatlah curam dan kurang lebar persebarannya berarti variasi biaya administrasi sangat kecil, sehingga penetapan harga lumayan bagus

California agak rendah dan agak lebar, variasinya tinggi Newyork skewness positif, berarti memang karena kota ekonomi, maka harga administrasi lumayan mahal memang wajar

marketing spend

boxplot

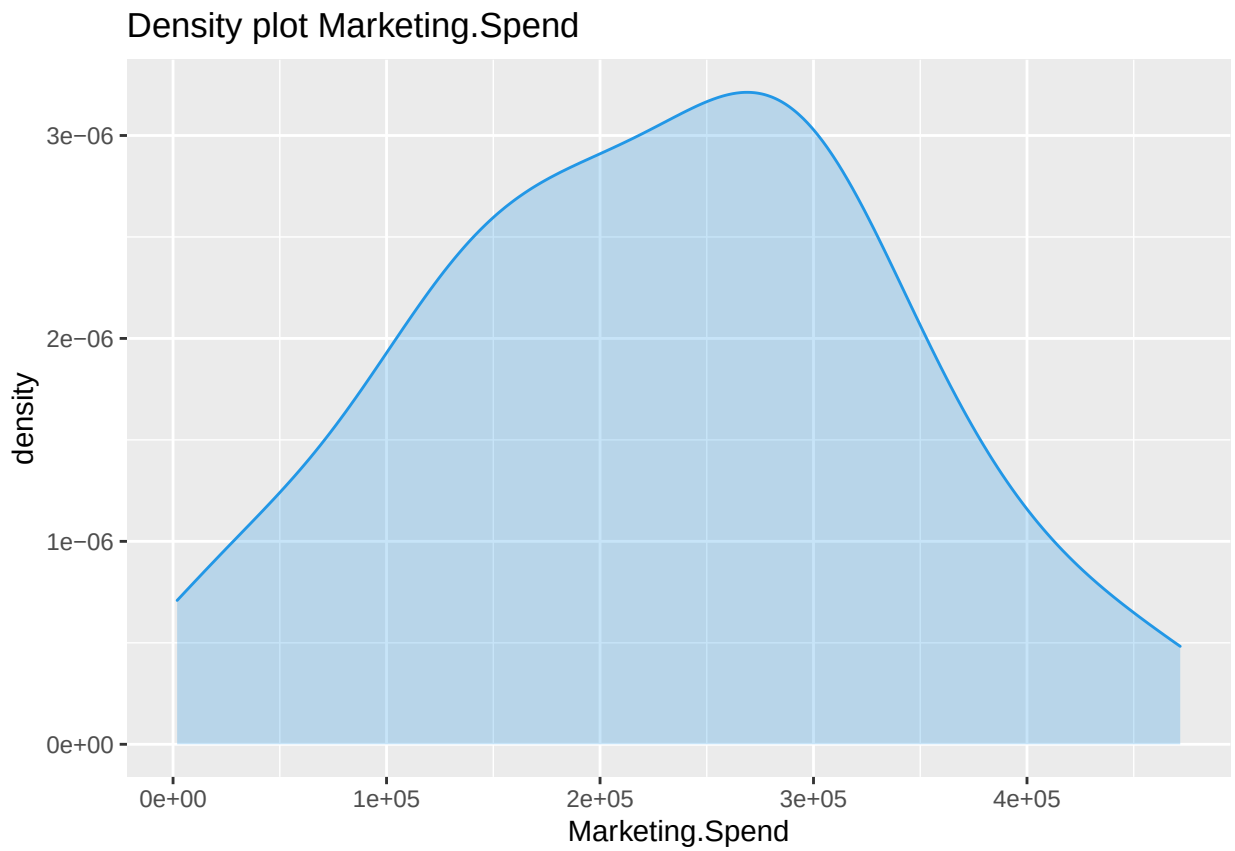
```
ggplot(dataku,aes(x=State,y=Marketing.Spend,fill=State)) +
  ggtitle("BoxPlot pada Marketing.Spend per State ")+
  geom_boxplot()
```



Markeing pada daerah Newyork lebih tinggi dibanding yang lain dan variatif, mungkin dikarenakan persaingan yang sangat berat karena dibutuh pengiklanan dan promosi yang besar besaran. Untuk California,mungkin disini ada perusahaan yang tujuan pemasarannya yang sangat luas dibanding yang lain, maka dari itu butuh iklan yang lebih besar daripada yang lain

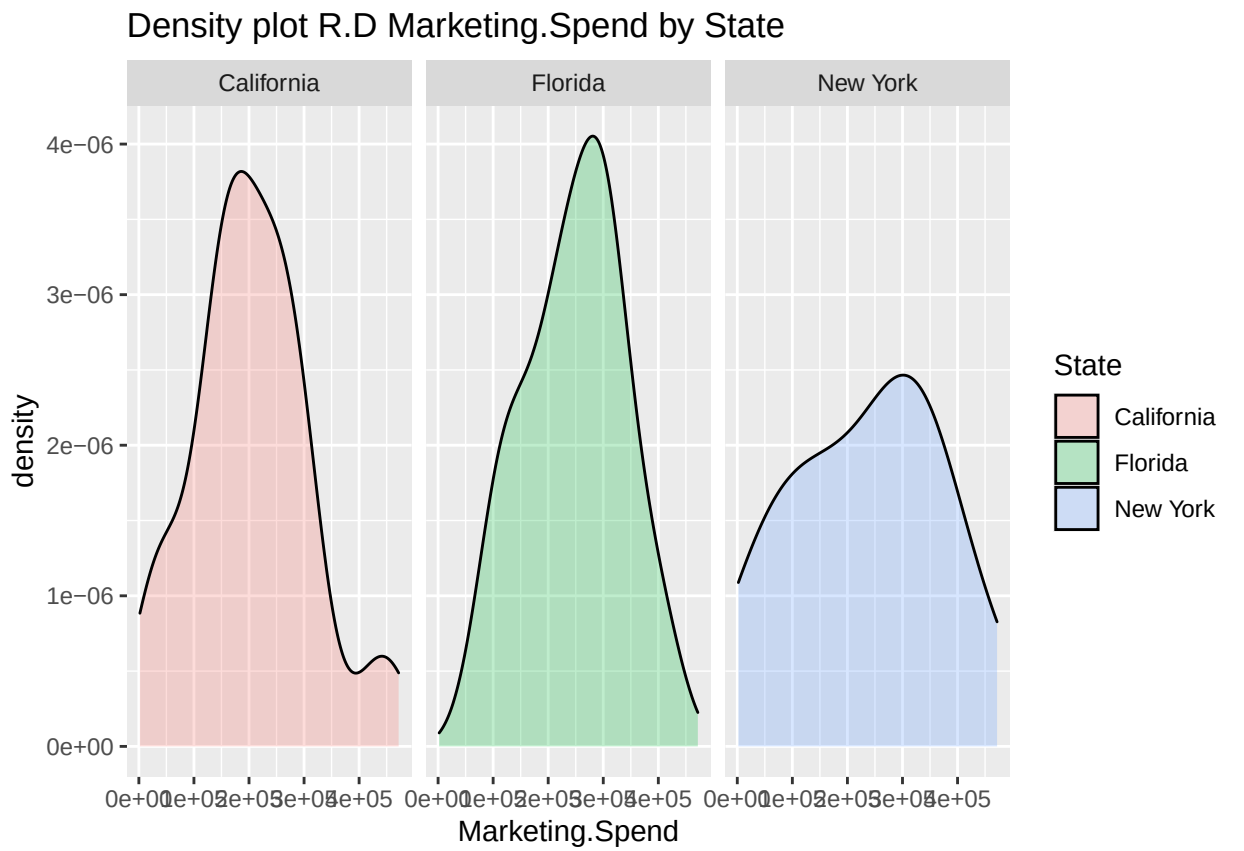
Density plot

```
ggplot(dataku,aes(x=Marketing.Spend)) +
  geom_density(color=4,fill=4,alpha=0.25) +
  ggtitle("Density plot Marketing.Spend")
```



Disini mendekati normal ya, maka lumayan baik tersebar merata. persaingannya sewajarnya

```
ggplot(dataku,aes(x=Marketing.Spend)) +
  geom_density(mapping=aes(fill=State),alpha=0.25) +
  ggtitle("Density plot R.D Marketing.Spend by State") +
  facet_wrap(~ State, ncol = 3)
```

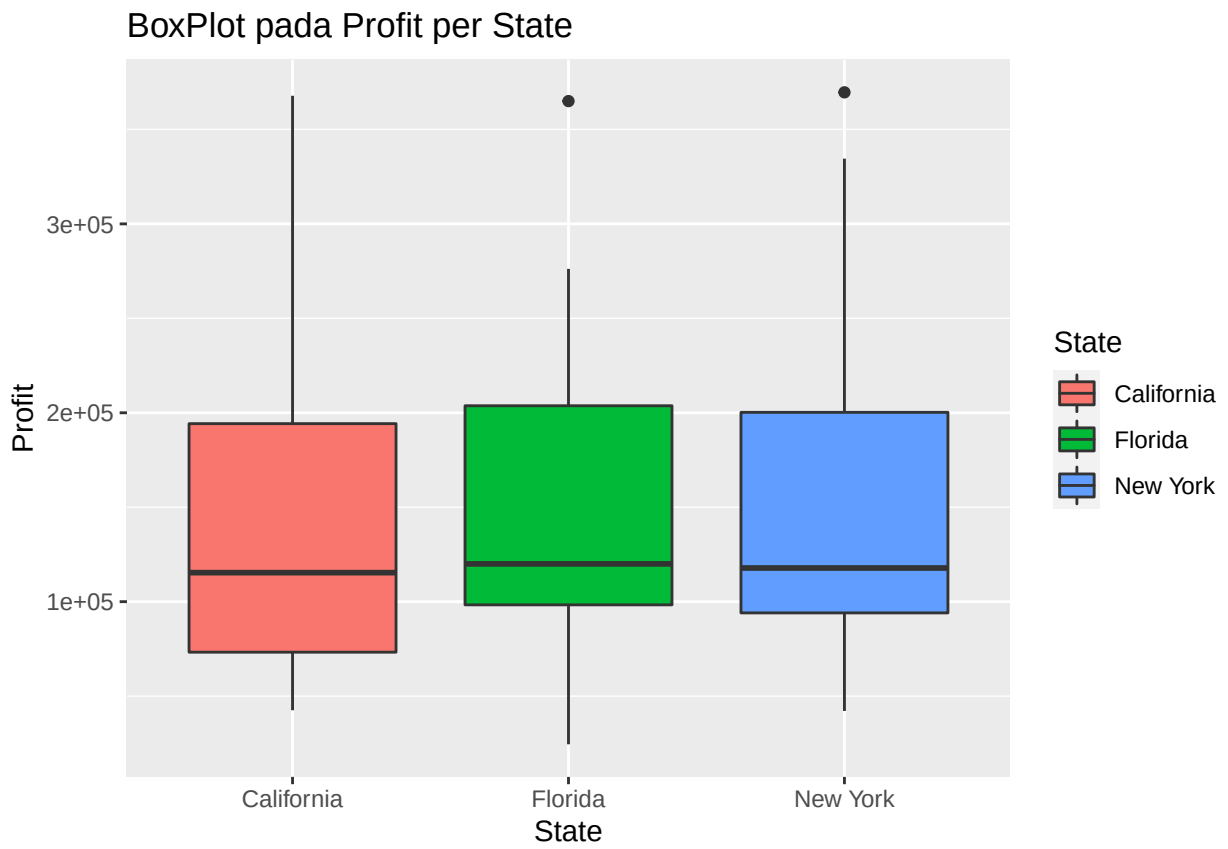


Disini di state masing masing normal, mungkin hanya persebaran newyork yang lebih lebar dikarenakan persaingan ini membutuhkan iklan yang lebih banyak

Profit

Boxplot

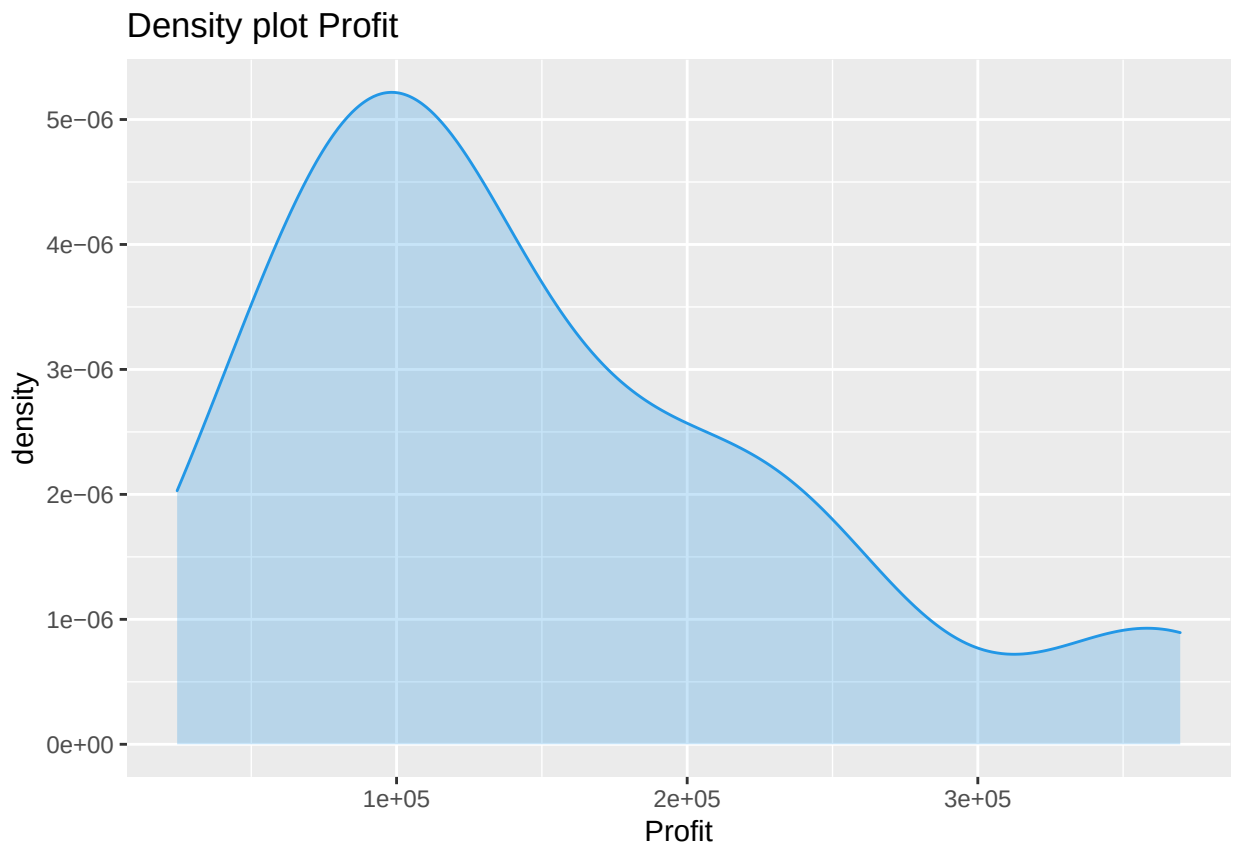
```
ggplot(dataku,aes(x=State,y=Profit,fill=State)) +
  ggtitle("BoxPlot pada Profit per State ")+
  geom_boxplot()
```



Persebarannya hampir sama antar Florida dan Newyork, meskipun Newyork lebih panjang, ternyata di florida dan newyork ada perusahaan besar yang menguasai di negara florida atau new york

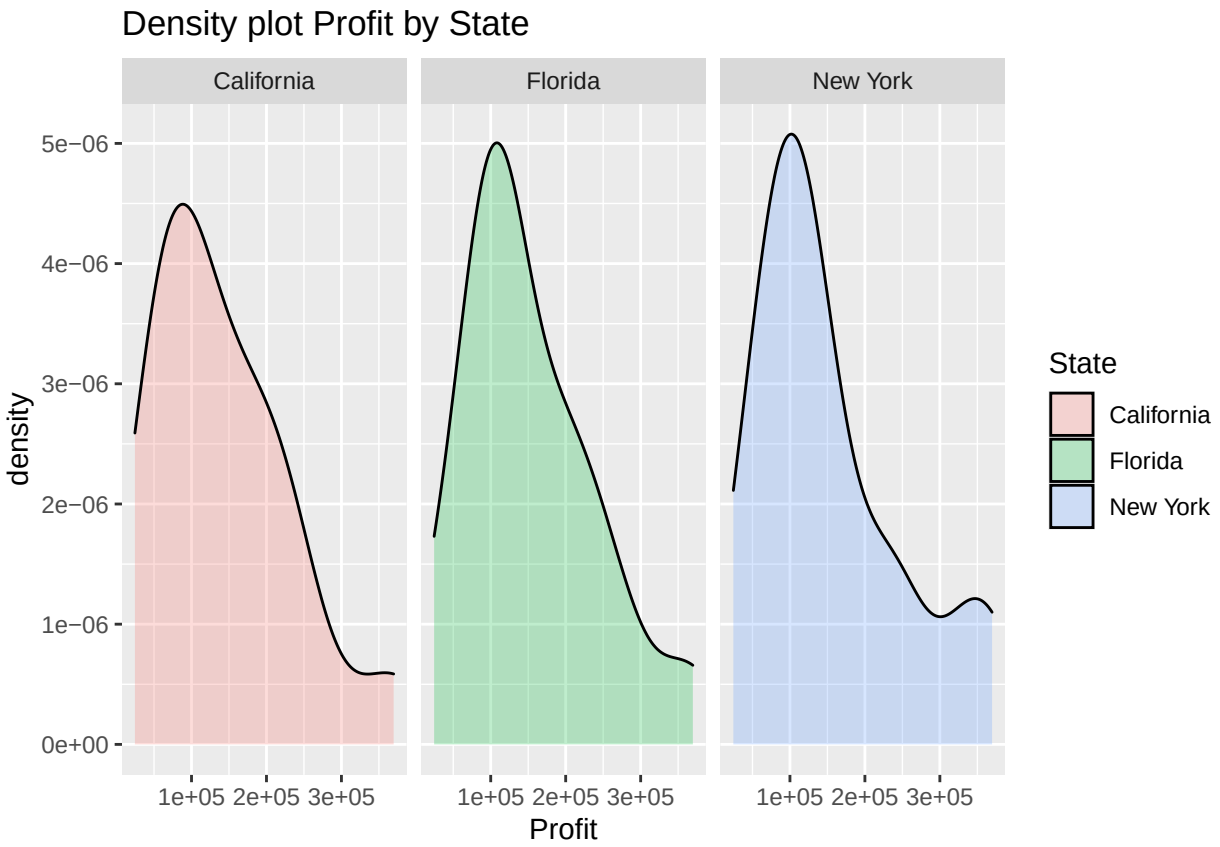
Density plot

```
ggplot(dataku,aes(x=Profit)) +
  geom_density(color=4,fill=4,alpha=0.25) +
  ggtitle("Density plot Profit")
```



Profit ini Skewness , berarti kebanyakan perusahaan sangatlah berjuang dalam mendapatkan keuntungan meskipun untungnya tersebut kecil, dan ada kemungkinan ada perusahaan raksasa yang menguasai pasar

```
ggplot(dataku,aes(x=Profit)) +
  geom_density(mapping=aes(fill=State),alpha=0.25) +
  ggtitle("Density plot Profit by State") +
  facet_wrap(~ State, ncol = 3)
```

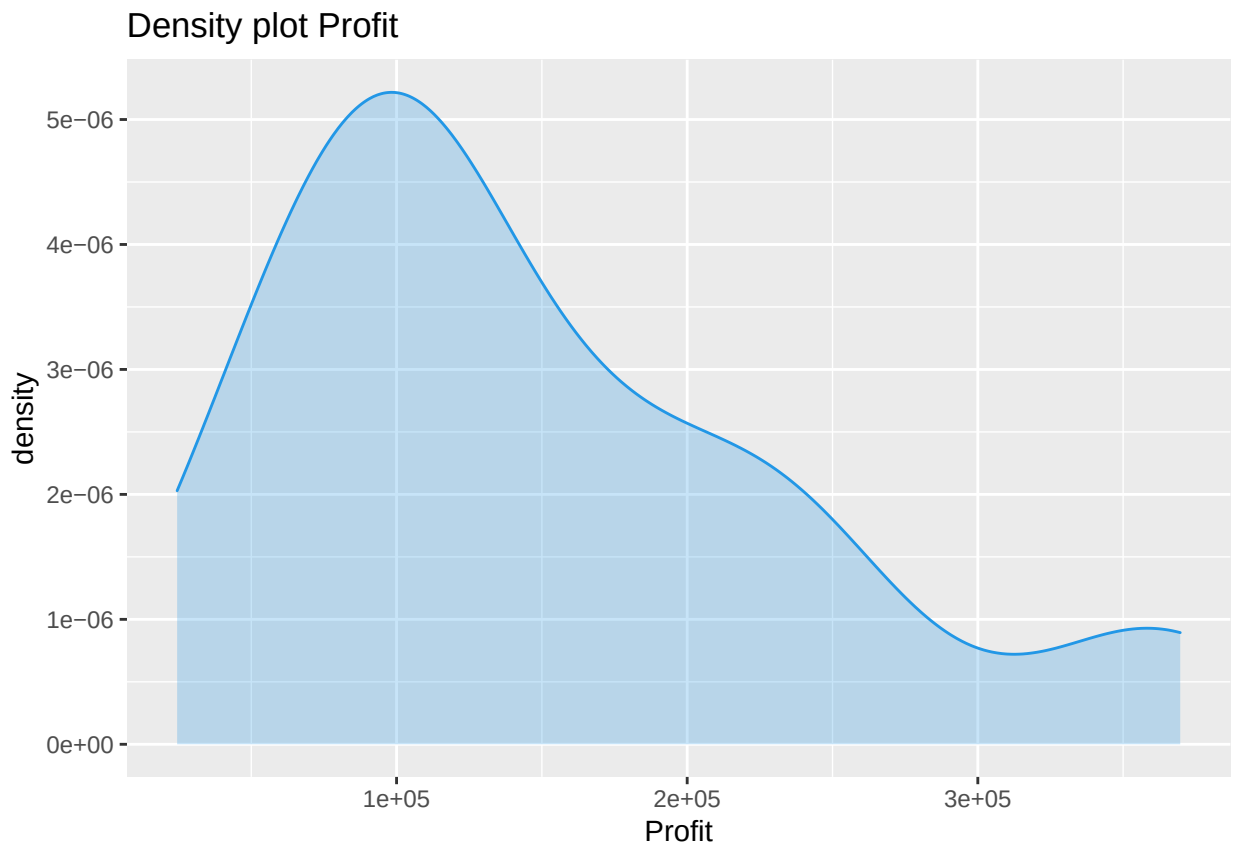


Sudah dibuktikan, bahwa keuntungan ini dikuasai Perusahaan raksasa di seluruh 3 kota meskipun perusahaan tersebut berbeda beda, maupun ada

Soal Nomor 2

Density Plot profit

```
ggplot(dataku,aes(x=Profit)) +
  geom_density(color=4,fill=4,alpha=0.25) +
  ggtitle("Density plot Profit")
```



Disini bukan Normal yah jika kita lihat secara Subjektif, dikarenkaan bisa dilihat sendiri, pertama Skewness positif atau bisa kita katakan berarti tidak simertris,berarti tidak normal dan juga tidak berbentuk seperti bell, maka tidak normal.

Untuk lebih yakin,maka kita uji dengan shapiro test karena lebih baik di test dibanding hanya dari pandangan sahaja

```
shapiro.test(dataku$Profit)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dataku$Profit
## W = 0.90149, p-value = 0.0009204
```

Kan didalam Shapiro itu, jika null Hipotesis maka normal, alternativenya , tidak normal.

maka, dikarenakan p-value < 0.05, –> tolak H0

maka Profit tidak Normal ya

dasar shapiro wilk itu seperti Artikel ini

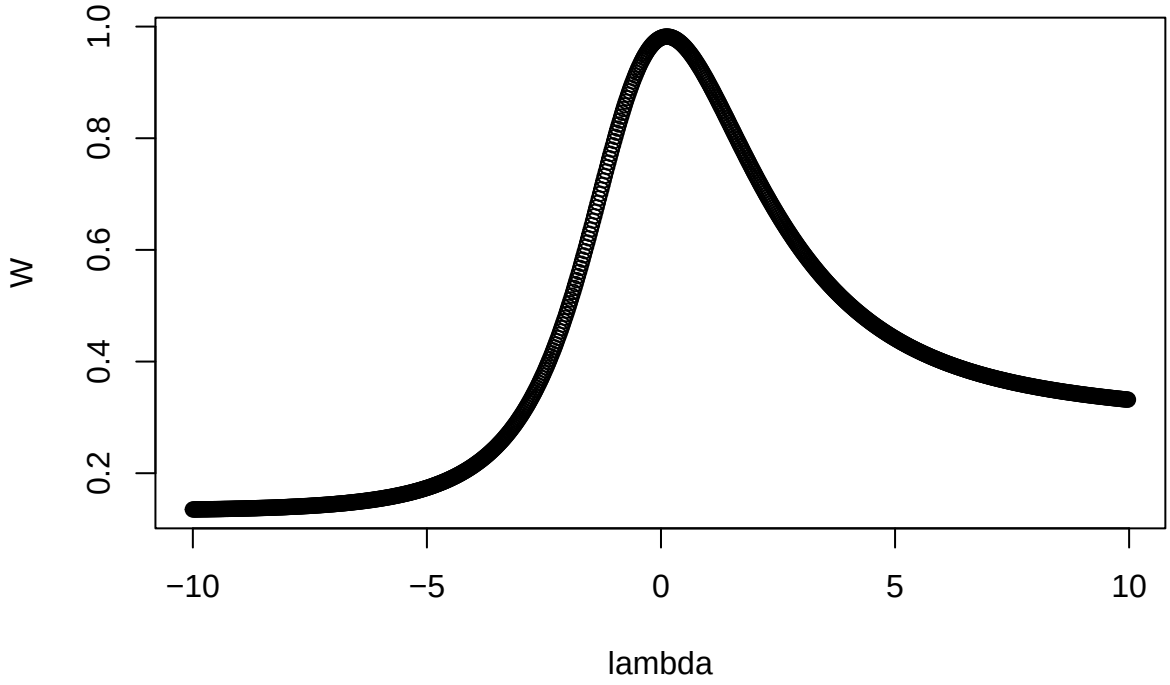
ini seperti menghitung selisih data asli dan data normal, kemudian dijumlahkan semuanya

Soal Nomor 3

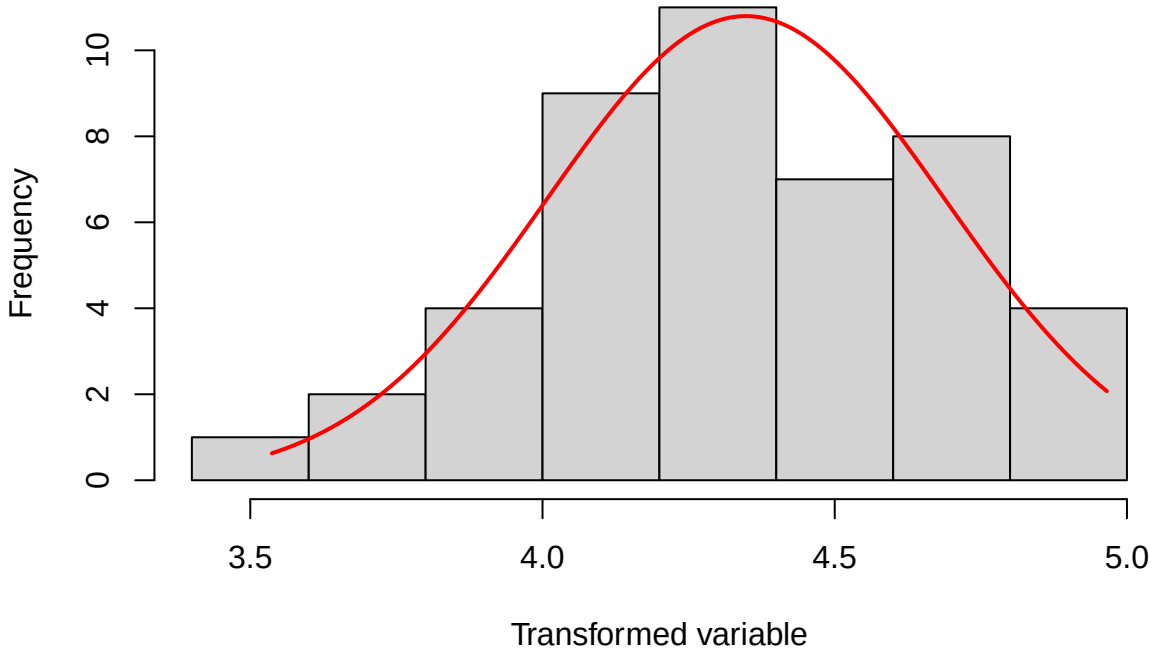
TransformTukey

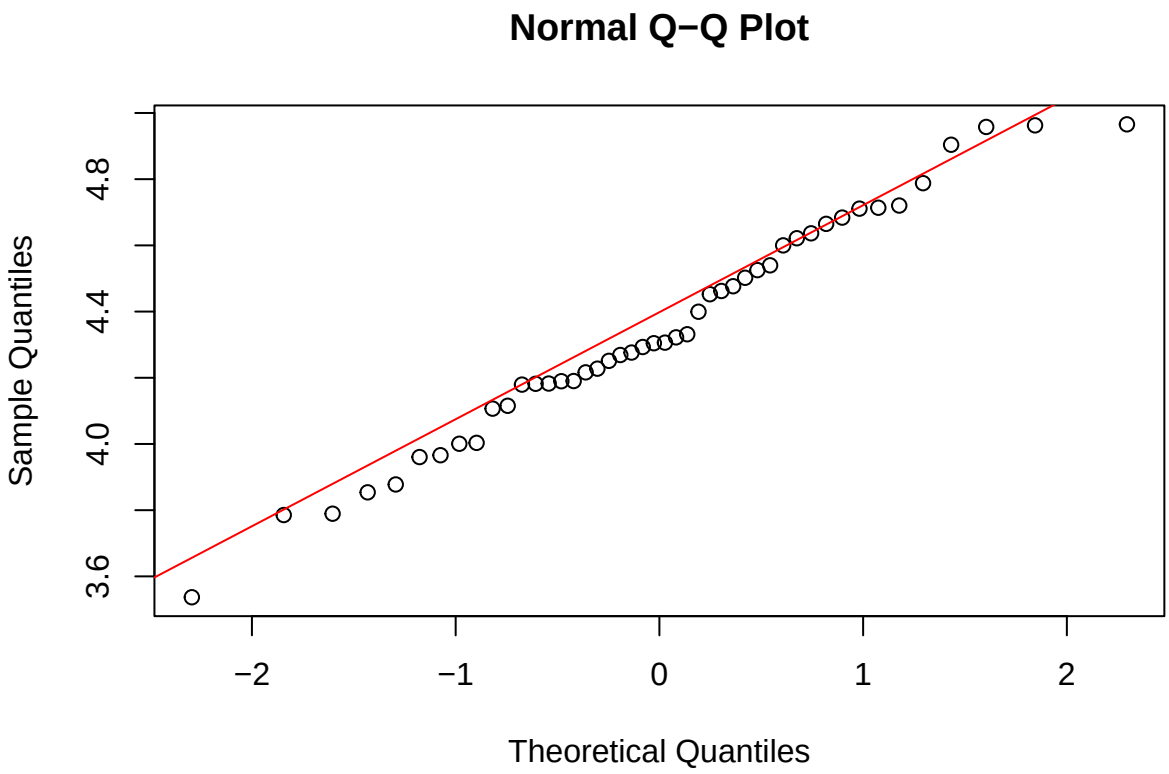
Dikarenakan pada test shapiro tersebut telah terbukti profit tidak normal maka lebih baik dilakukan tranformasi

```
profit_tukey <- transformTukey(dataku$Profit)
```




```
##
##     lambda      W Shapiro.p.value
## 406  0.125 0.9821          0.6917
##
## if (lambda >  0){TRANS = x ^ lambda}
## if (lambda == 0){TRANS = log(x)}
## if (lambda <  0){TRANS = -1 * x ^ lambda}
```


Frequency
0
2
4
6
8
10
3.5
4.0
4.5
5.0
Transformed variable



Dari hasil tersebut, ada plot W Vs lambda, itu yang W itu adalah hasil perhitungan dari shapiro test ya dan lambda adalah lambda pada rumus Tukey nya

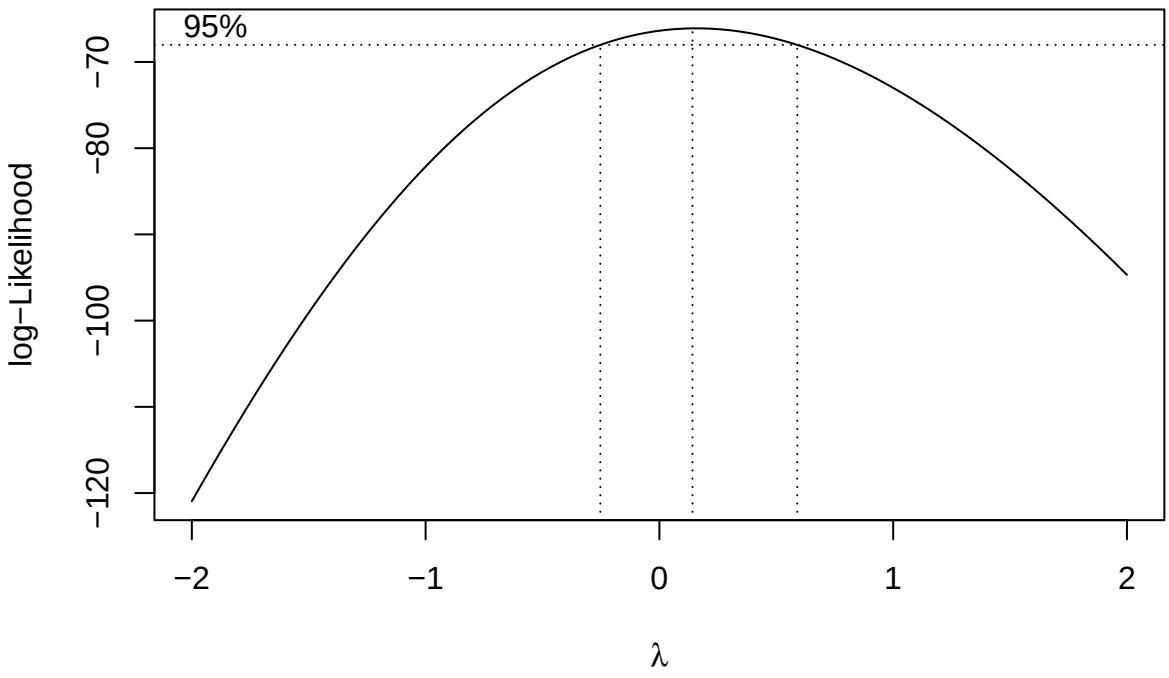
ini site untuk rumus lambdanya

Dikarenakan lambdanya adalah 0.125 maka hasilnya adalah, dipangkatkan 0.125 berarti data tersebut di akar pangkat 8 agar lebih normal dan dengan test shapiro juga sudah ditest normal ya.

oh iya jika kurang dari 1 lamdanya berarti itu dalam kondisis skewness positif (menjulur ke kanan) kalau lebih dari lambdanya 1 berarti kondisi tersebut menjulur kekiri (skewness negatif)

Transformasi boxcox

```
bc_pr <- boxcox(dataku$Profit ~ 1)
```




```
bc_lambda <- bc_pr$x[which.max(bc_pr$y)]
bc_profit <- (dataku$Profit^bc_lambda - 1)/bc_lambda
```

untuk teorinya ada di sini ya

```
# Disini dengan menggunakan shapiro test, maka bisa ditentokan transform tukey
# dengan boxcox

aa <- shapiro.test(profit_tukey)
bb <- shapiro.test(bc_profit)

print(paste("tukey",aa$statistic))
```

```
## [1] "tukey 0.982071240851543"
```

```
print(paste("boxcox",bb$statistic))
```

```
## [1] "boxcox 0.982034471414615"
```

```
aa$statistic > bb$statistic
```

```
##    W
## TRUE
```

Dikarenakan lebih besar adalah tukey, maka yang digunakan adalah tukey

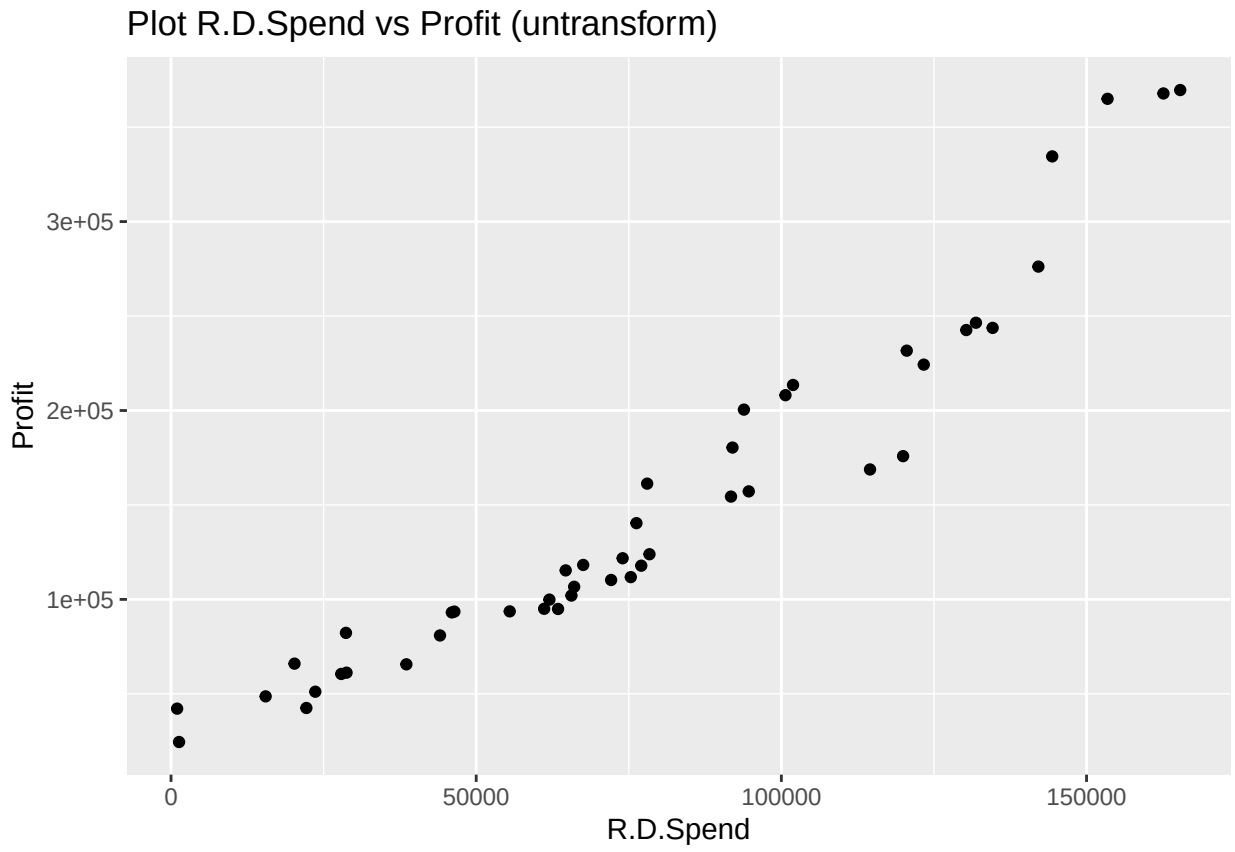
```
dataku_transform = dataku
dataku_transform$Profit <- profit_tukey
```

Soal Nomor 4

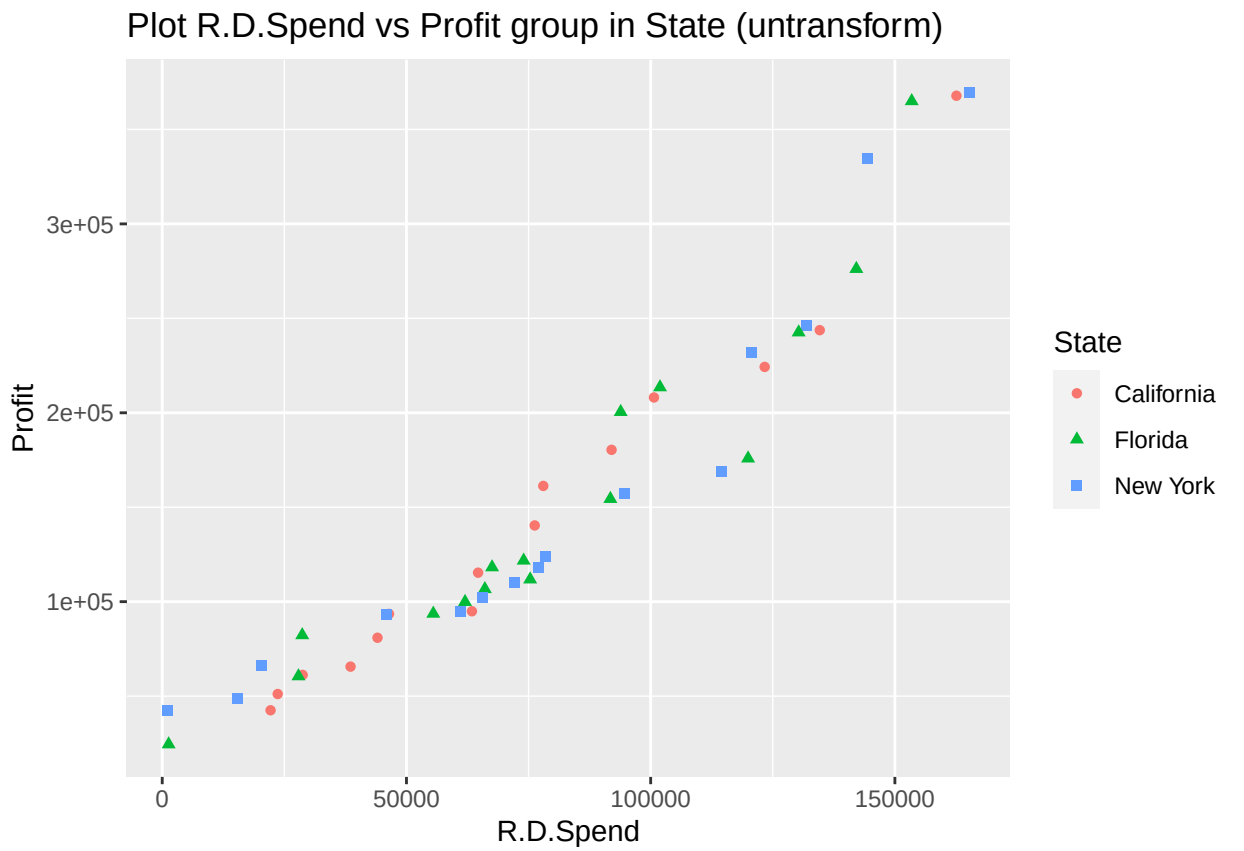
Dikarenakan Saya berasumsi bahwa dalam dataset tersebut, yang dicari adalah profit maka saya akan plot selalu terhadap profit

R.D.Spend_untransformed

```
ggplot(dataku,aes(x=R.D.Spend,y=Profit)) +
  ggtitle("Plot R.D.Spend vs Profit (untransform)") +
  geom_point()
```

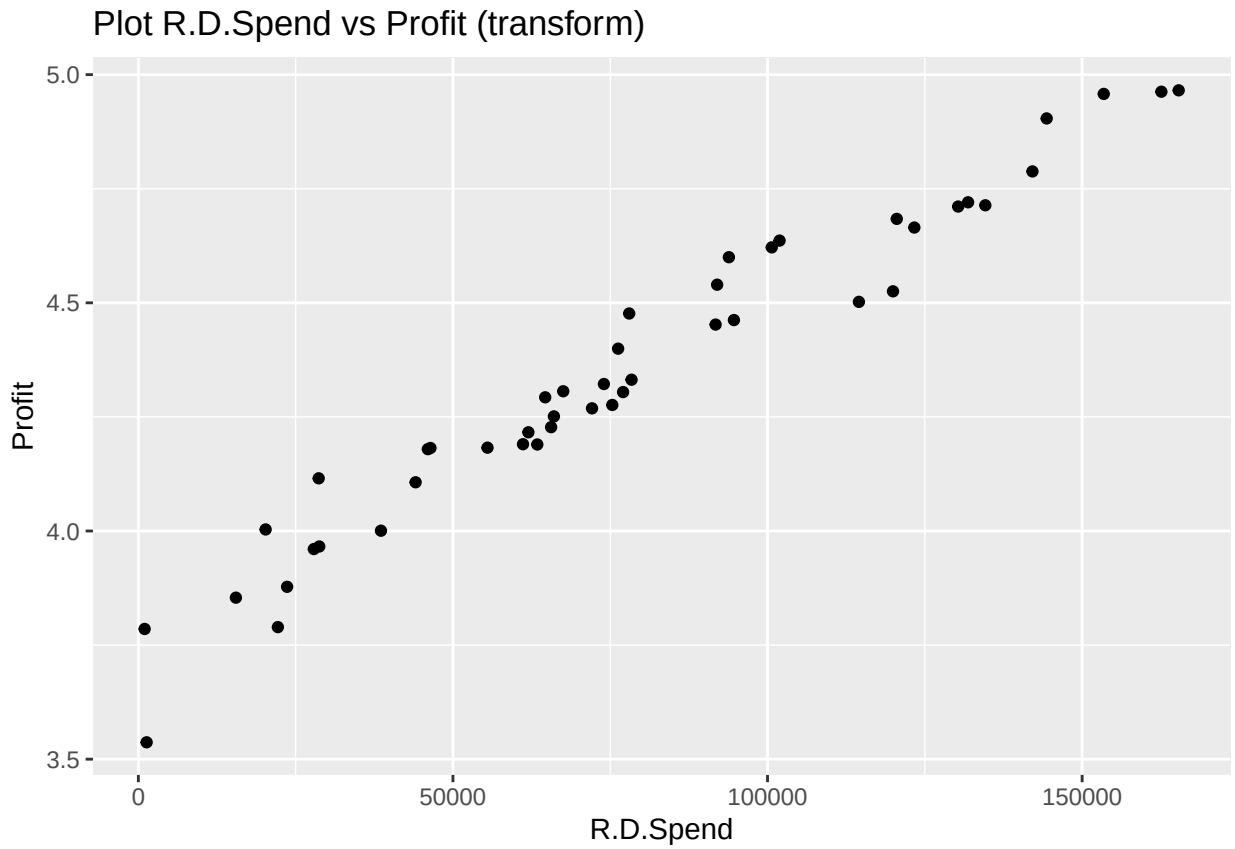



```
ggplot(dataku,aes(x=R.D.Spend,y=Profit,
                  shape=State,
                  color=State)) +
  ggtitle("Plot R.D.Spend vs Profit group in State (untransform)") +
  geom_point()
```

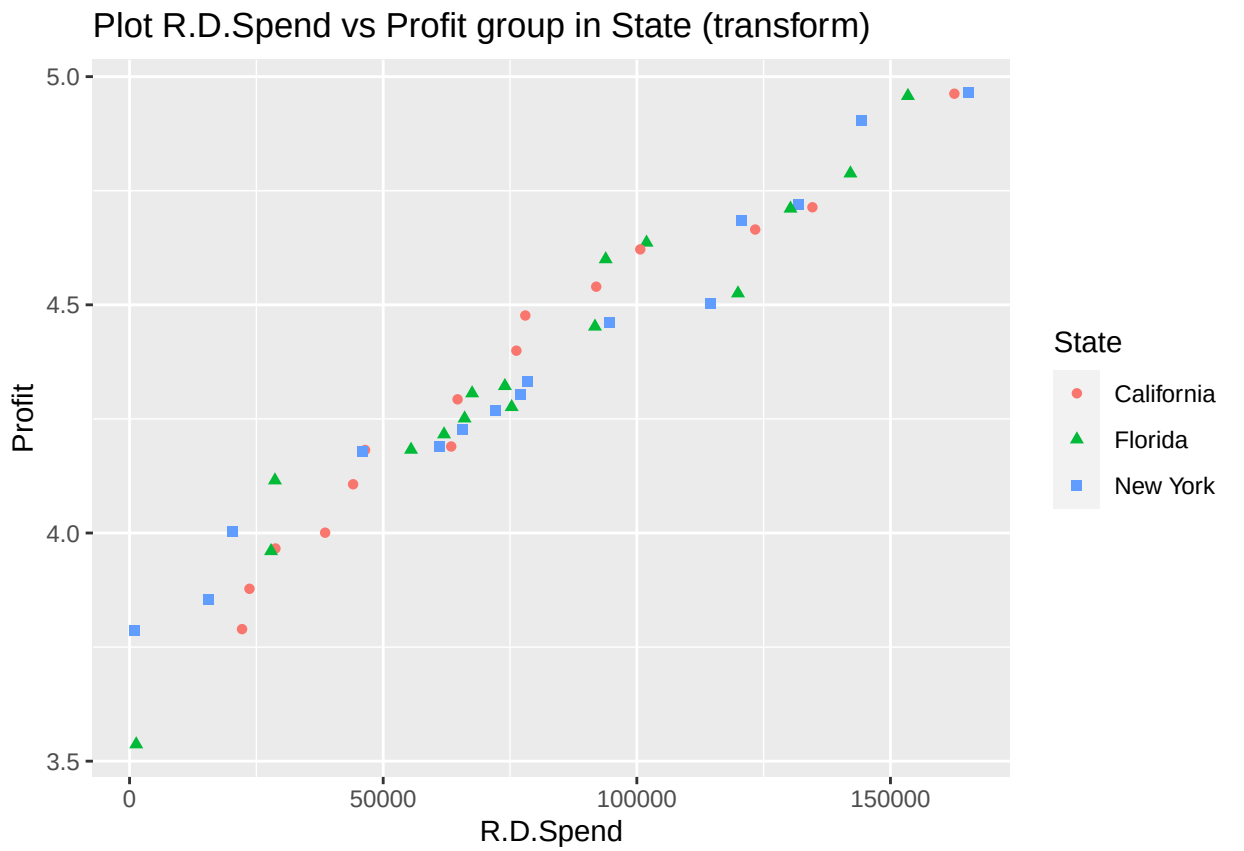


Di plot 1 , disini ada kecenderungan cekung ke atas yaa , mungkin harus di transform Di plot 2, disini tidak ada pengelompokan nilai tersebut ya

```
ggplot(dataku_transform,aes(x=R.D.Spend,y=Profit)) +
  ggtitle("Plot R.D.Spend vs Profit (transform)") +
  geom_point()
```



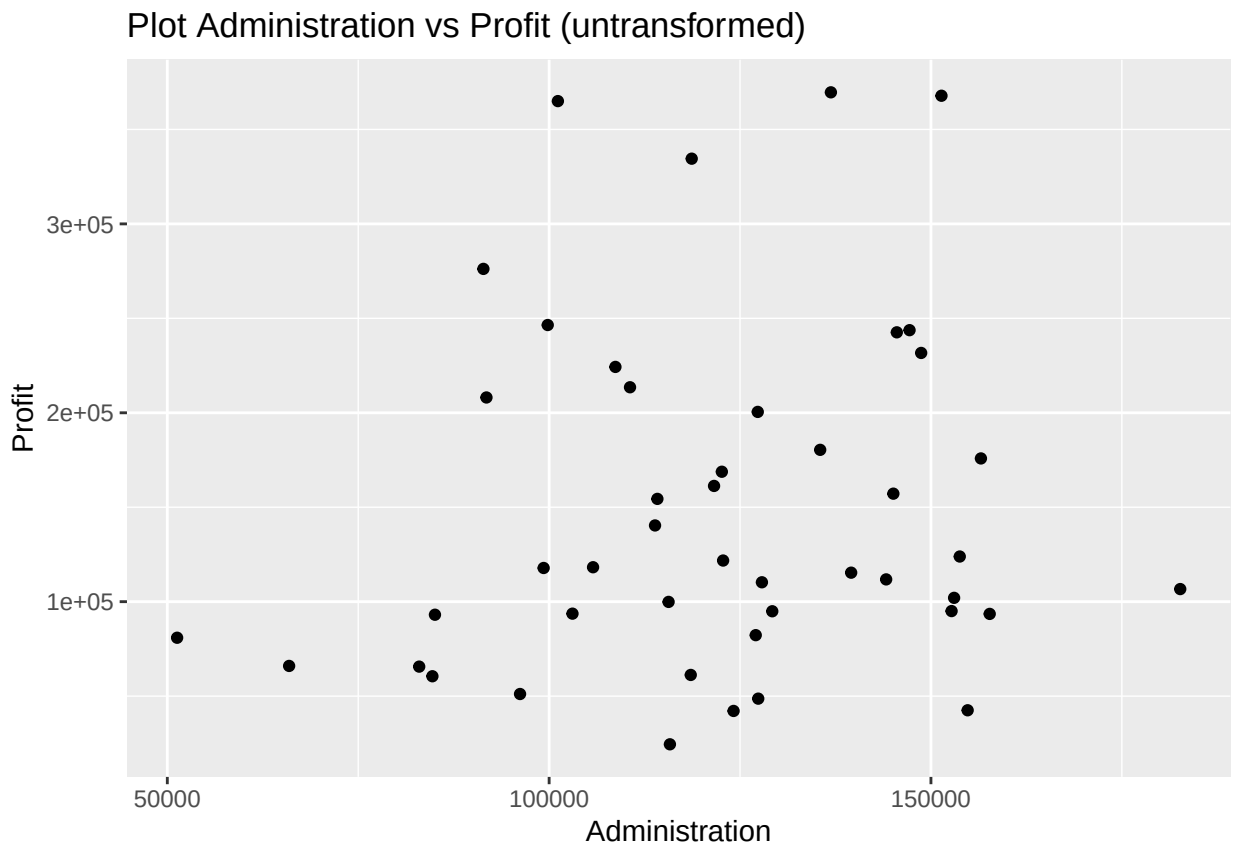

```
ggplot(dataku_transform,aes(x=R.D.Spend,y=Profit,
                  shape=State,
                  color=State)) +
  ggtitle("Plot R.D.Spend vs Profit group in State (transform)") +
  geom_point()
```



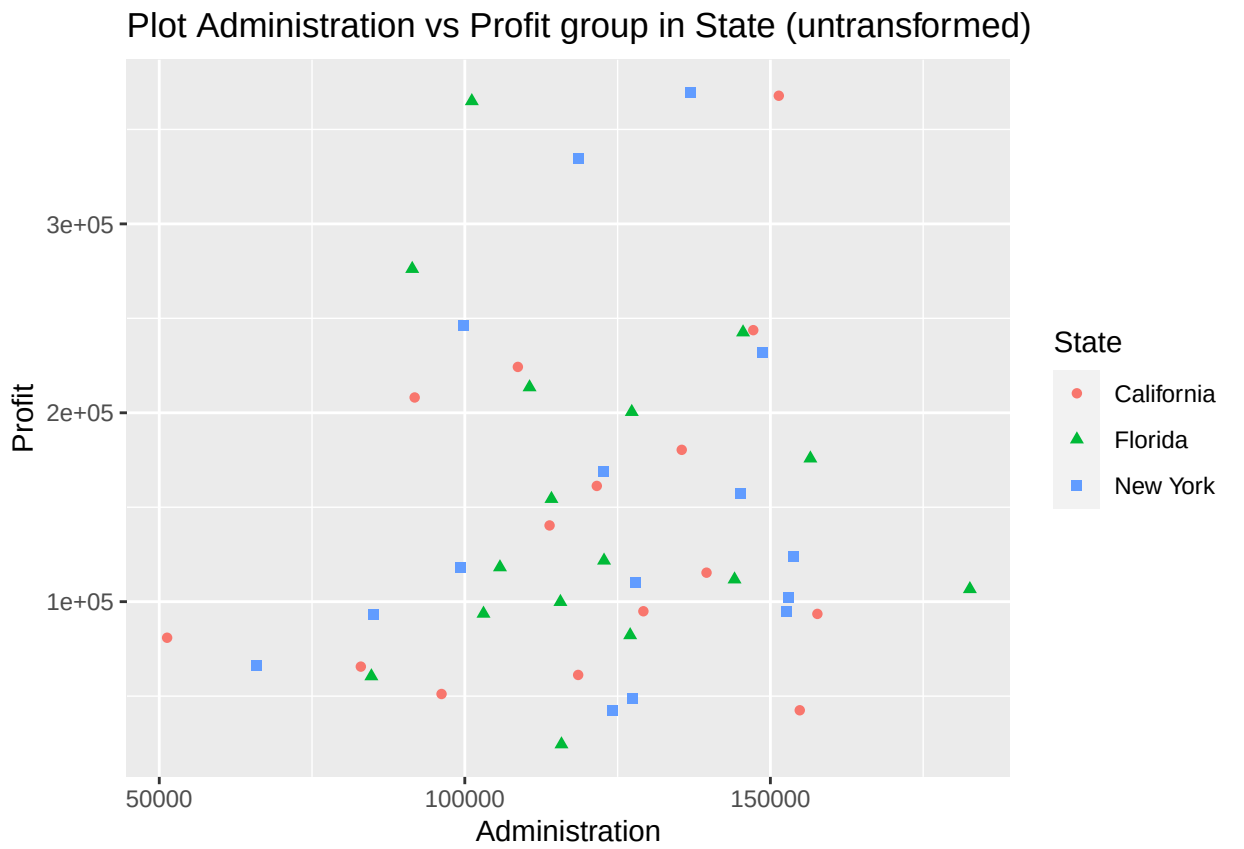
Di plot pertama, berkorelasi positif karena sudah jelas bergerak ke atas kanan , Di plot kedua, data antar statetnya sudah tercampur, sehingga tidak perlu adanya clustering Sehingga datanya tersebut jika di regresikan maka setiap group bergaris sama nilainya

Administration

```
ggplot(dataku,aes(x=Administration,y=Profit)) +
  ggtitle("Plot Administration vs Profit (untransformed)") +
  geom_point()
```



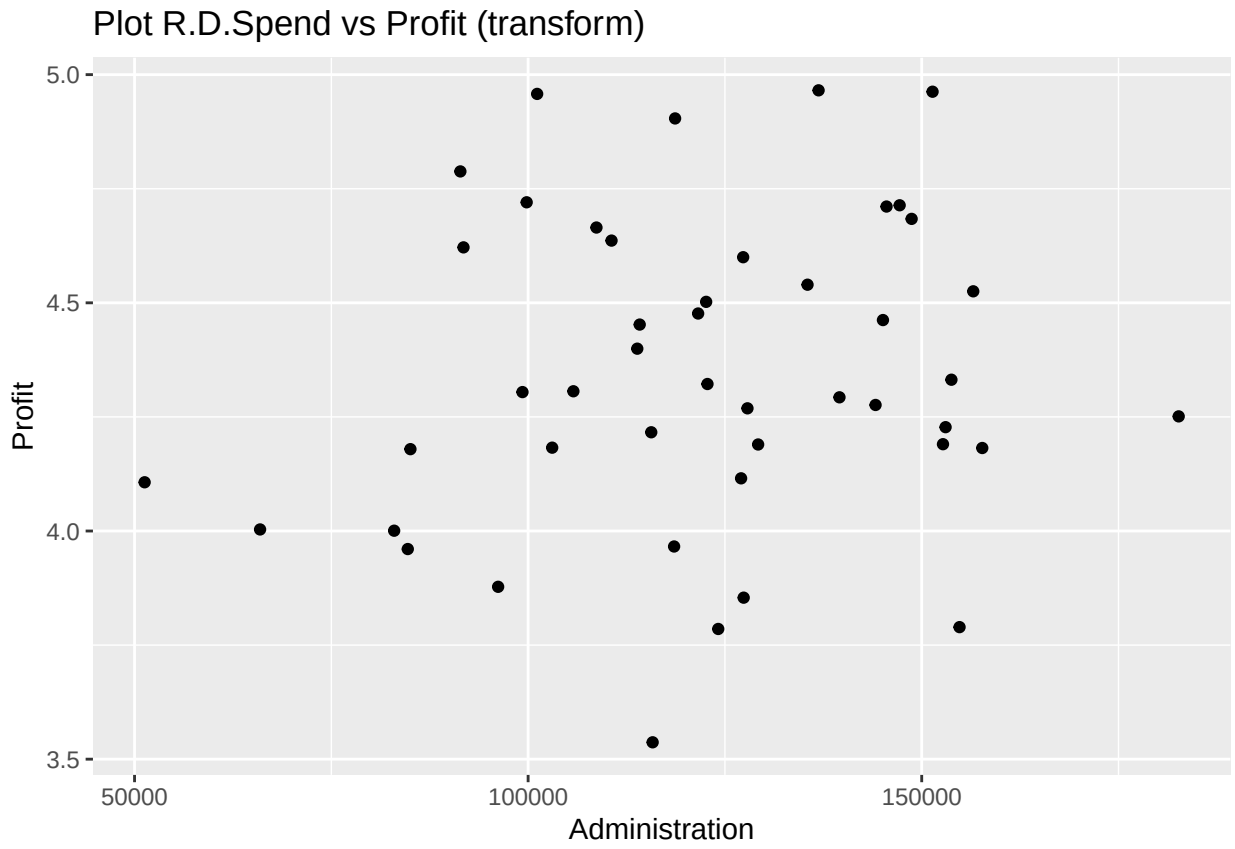

```
ggplot(dataku,aes(x=Administration,y=Profit,
                  shape=State,
                  color=State)) +
  ggtitle("Plot Administration vs Profit group in State (untransformed)") +
  geom_point()
```



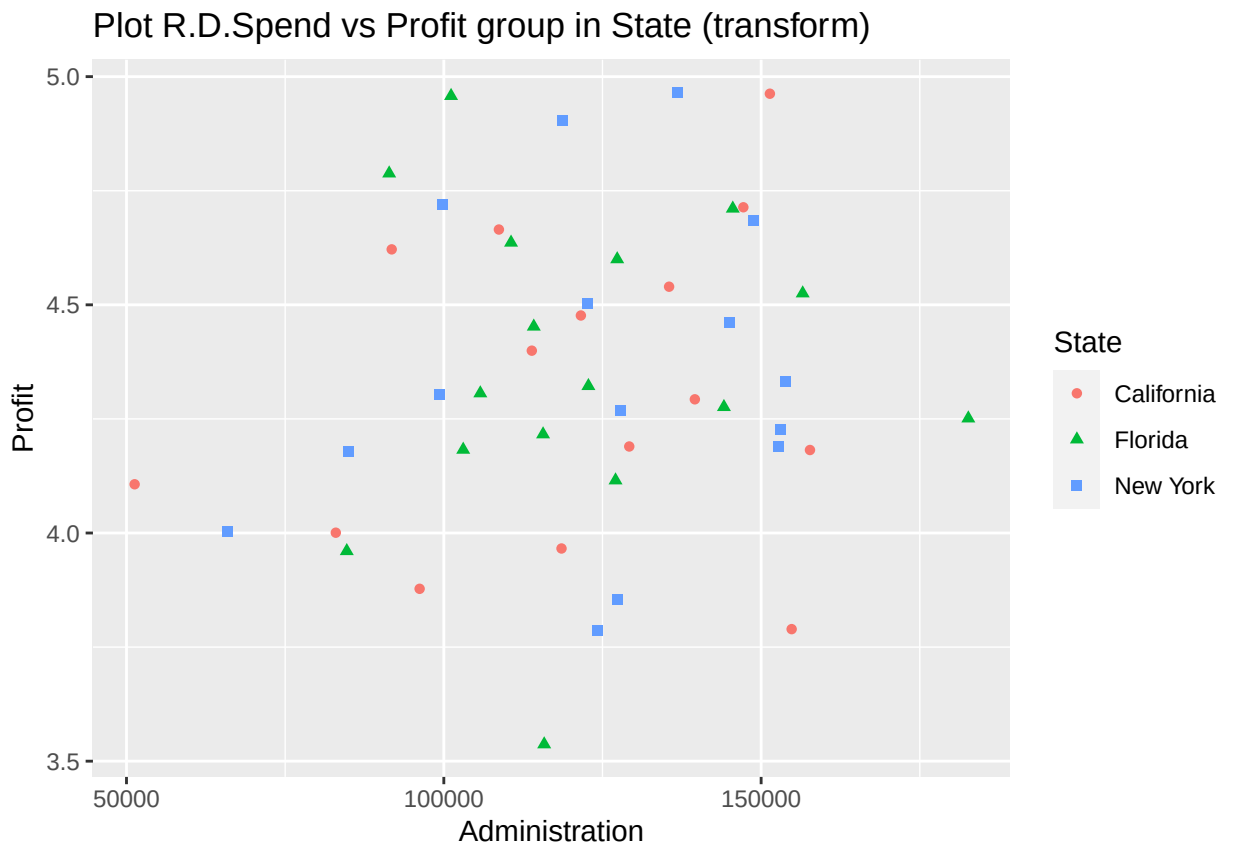
untuk plot pertama; Waduh, disini tidak terlihat berkorelasi baik negatif maupun positif ya, berarti dikarenakan secara logika administrasi dengan profit bisa dikatakan ada hubungannya, maka adminstrasi tidak mempengaruhi profit

untuk plot kedua, disini tidak ada pengelompokan data yang berarti

```
ggplot(dataku_transform,aes(x=Administration,y=Profit)) +
  ggtitle("Plot R.D.Spend vs Profit (transform)") +
  geom_point()
```




```
ggplot(dataku_transform,aes(x=Administration,y=Profit,
                  shape=State,
                  color=State)) +
  ggtitle("Plot R.D.Spend vs Profit group in State (transform)") +
  geom_point()
```

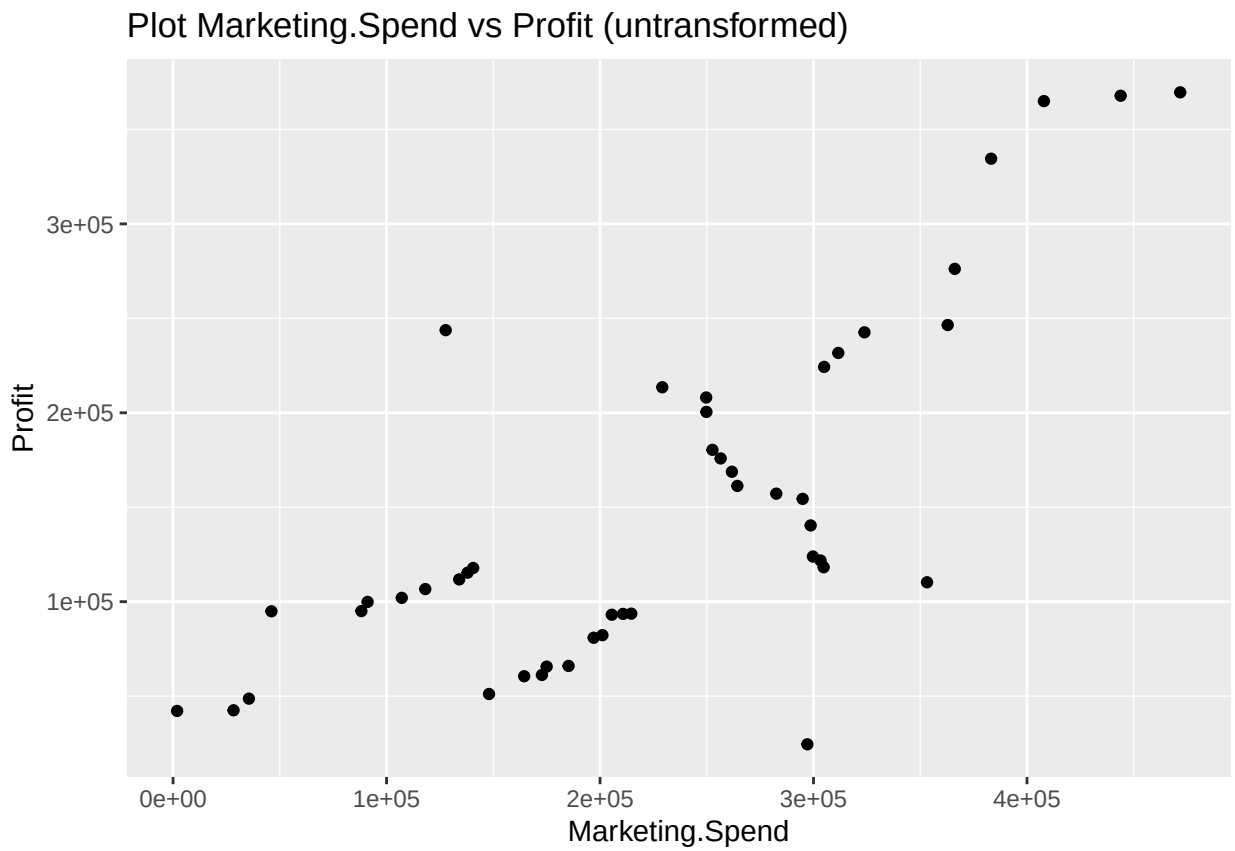


untuk plot pertama; Waduh, disini tidak terlihat berkorelasi baik negatif maupun positif ya, berarti dikarenakan secara logika administrasi dengan profit bisa dikatakan ada hubungannya, maka adminstrasi tidak mempengaruhi profit

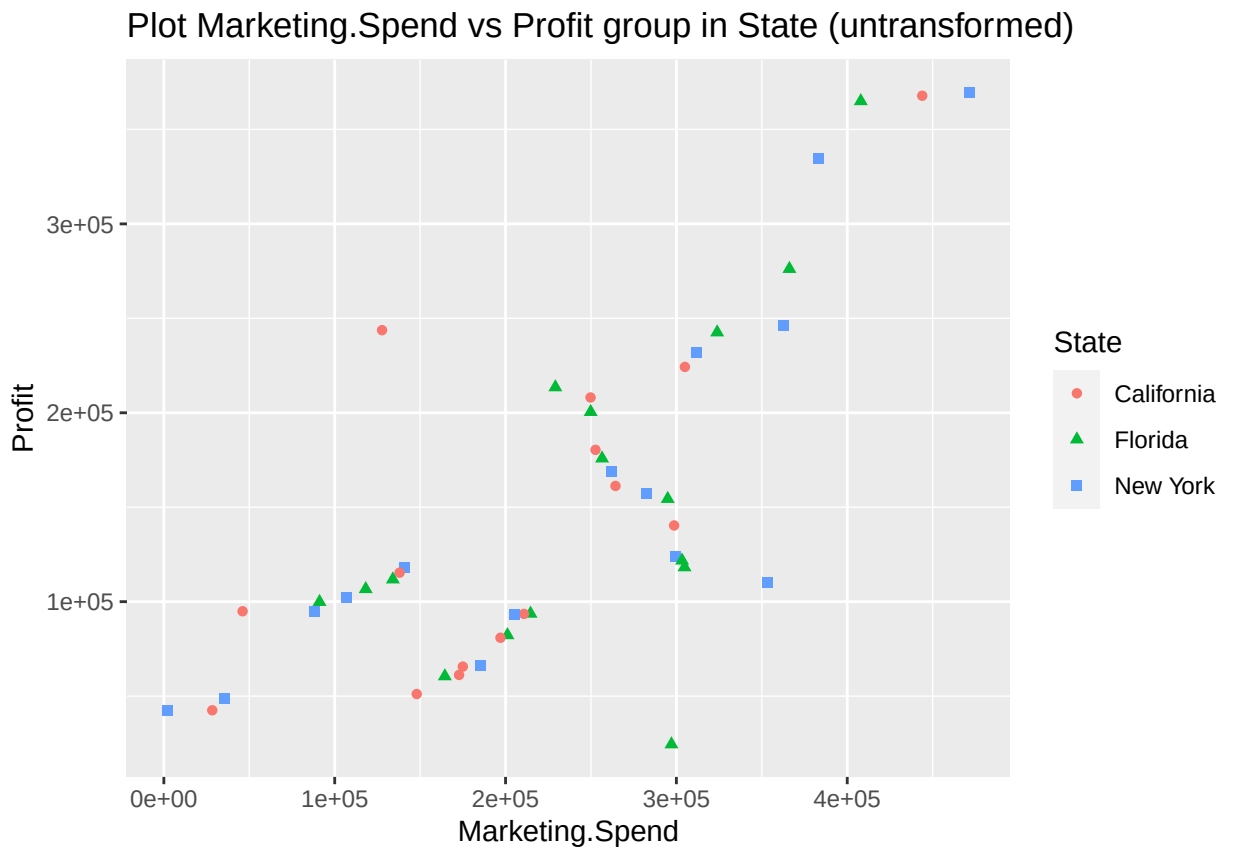
untuk plot kedua, disini tidak ada pengelompokan data yang berarti

Marketing Spend

```
ggplot(dataku,aes(x=Marketing.Spend,y=Profit)) +
  ggtitle("Plot Marketing.Spend vs Profit (untransformed)") +
  geom_point()
```



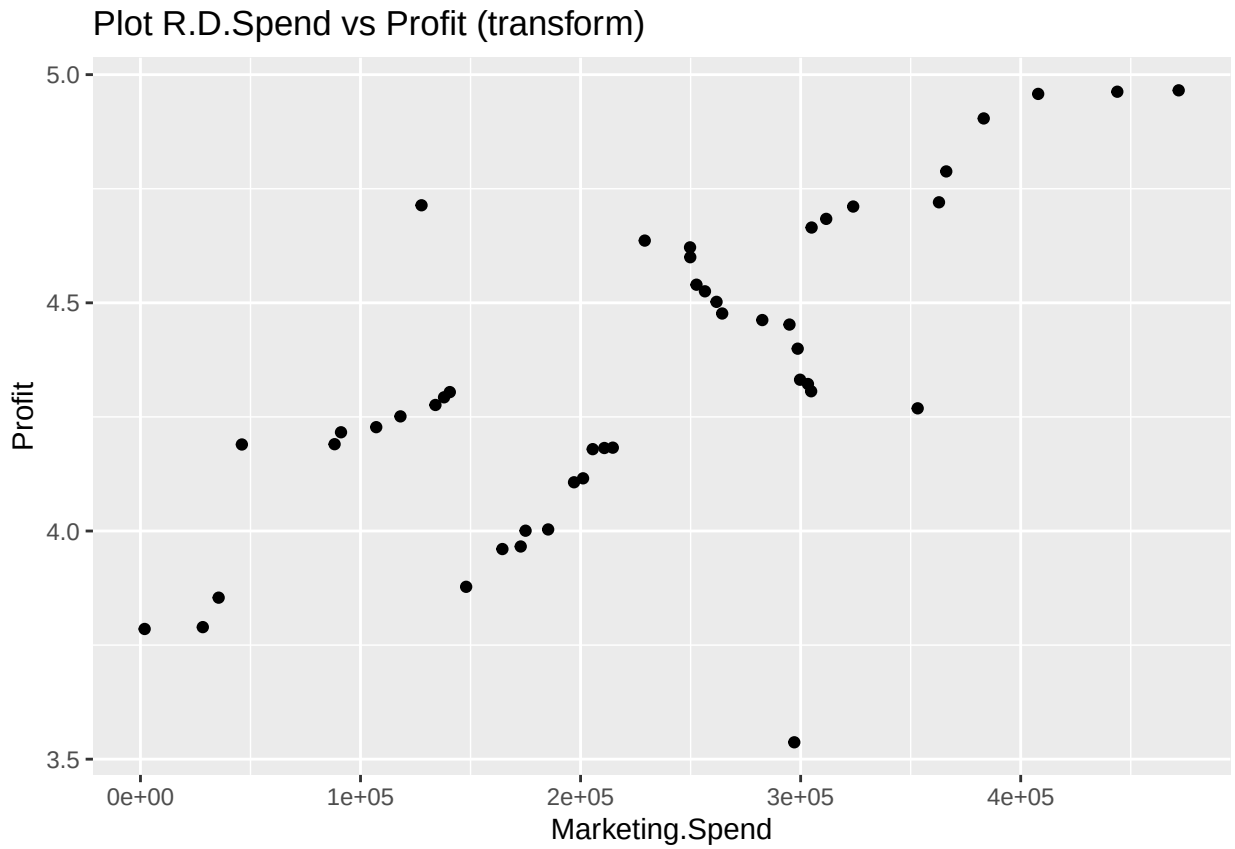

```
ggplot(dataku,aes(x=Marketing.Spend,y=Profit,
                  shape=State,
                  color=State)) +
  ggtitle("Plot Marketing.Spend vs Profit group in State (untransformed)") +
  geom_point()
```



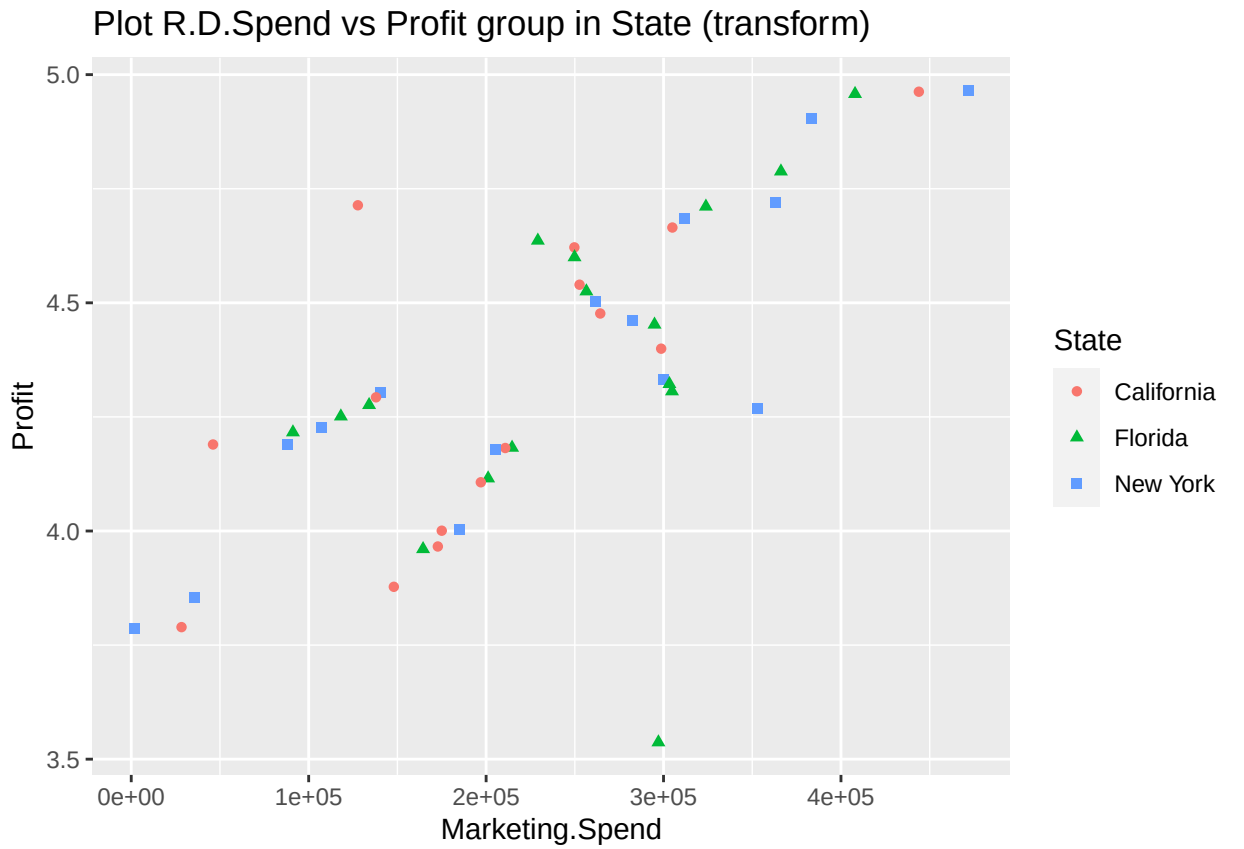
Untuk plot pertama ada suatu nilai yang berbentuk aneh tetapi jika diperhatikan maka berkoelasi positif ya

Unutk plot kedua, masih saja tercampur rata di statenya

```
ggplot(dataku_transform,aes(x=Marketing.Spend,y=Profit)) +
  ggtitle("Plot R.D.Spend vs Profit (transform)") +
  geom_point()
```




```
ggplot(dataku_transform,aes(x=Marketing.Spend,y=Profit,
                  shape=State,
                  color=State)) +
  ggtitle("Plot R.D.Spend vs Profit group in State (transform)") +
  geom_point()
```

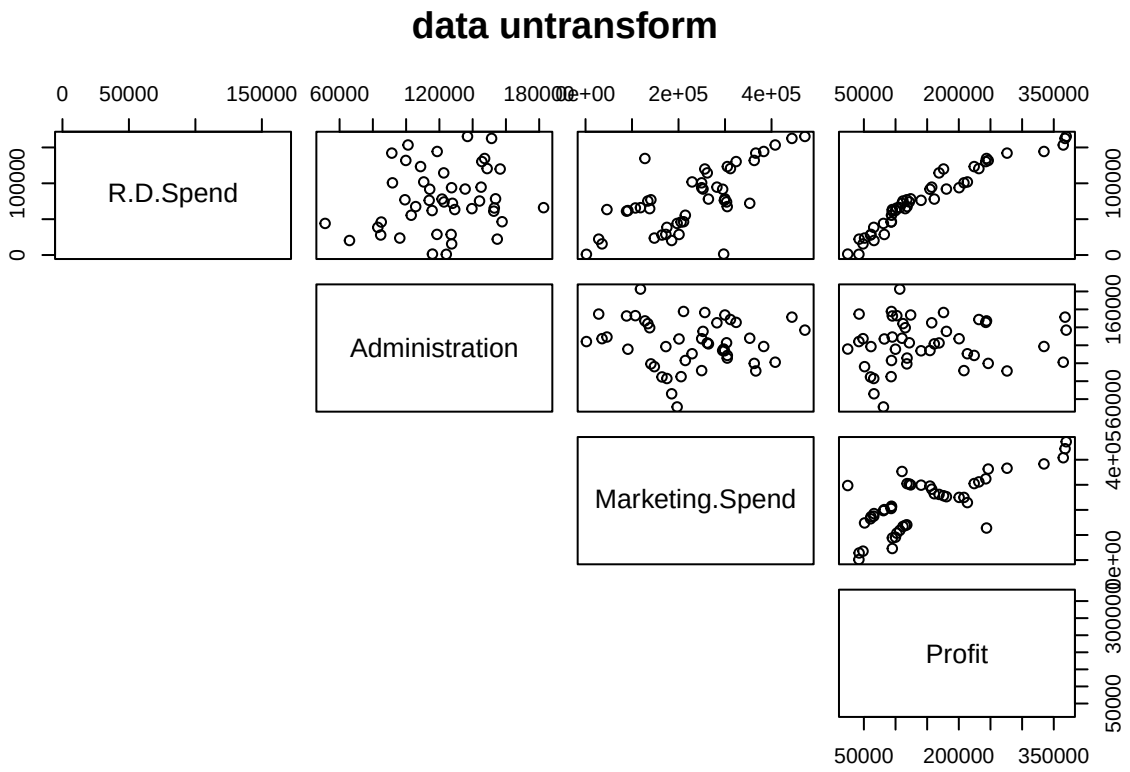


Untuk plot pertama ada suatu nilai yang berbentuk aneh tetapi jika diperhatikan maka berkoelasi positif ya

Unutk plot kedua, masih saja tercampur rata di statenya

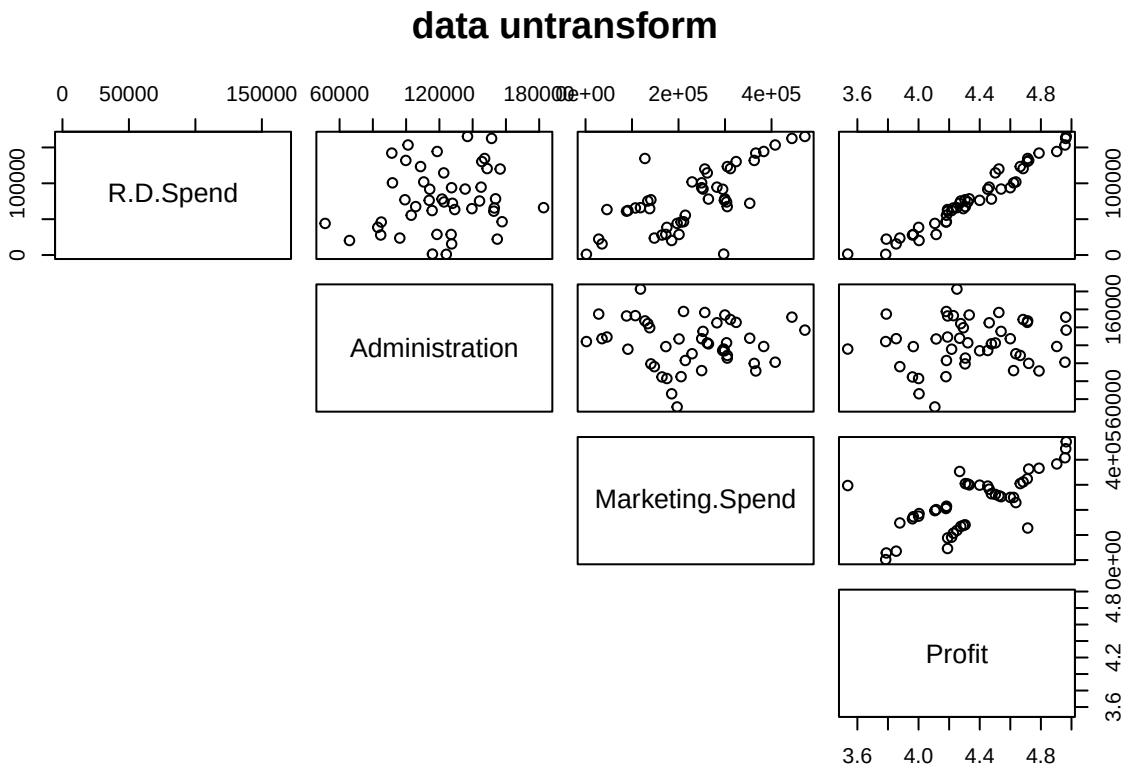
Nah, Untuk Marketing.Spend harus Ditransformasi ya dikarenakan bentuknya tidak terlihat jelas yaa

```
pairs(dataku[,num_kol],lower.panel = NULL,main="data untransform")
```



Disini , ada correlasi antara merketin spend dan R.D Spend, maka diperlukan tindakan agar tidak ada corealsi antar peubah (multicolinearity), maka harus diuji VIF, agar bisa baik yaah sehingga Profit bisa meningkat yaa

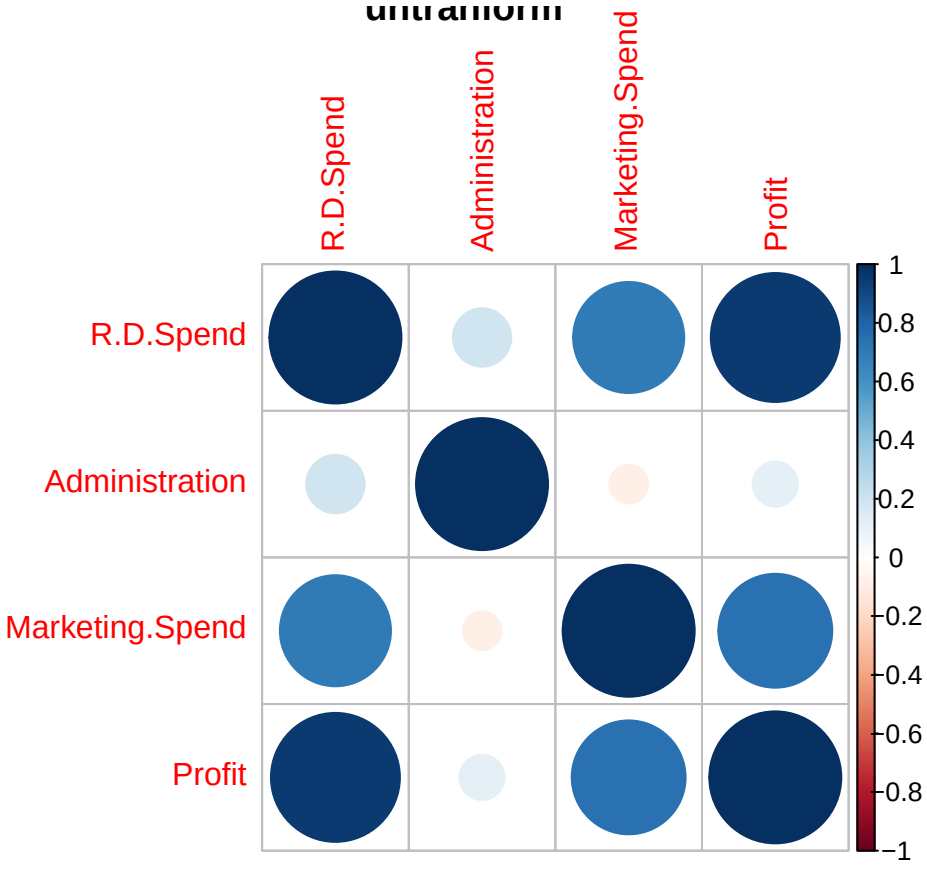
```
pairs(dataku_transform[,num_kol],lower.panel = NULL,main="data untransform")
```



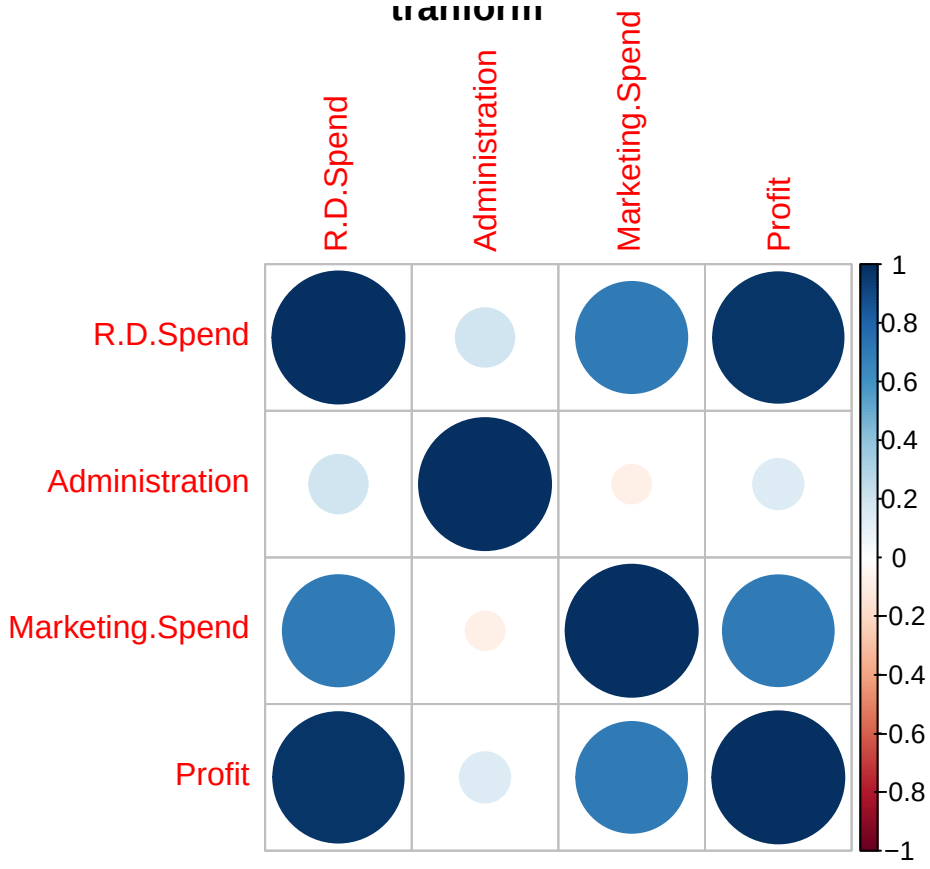
Disini , ada correlasi antara merketin spend dan R.D Spend, maka diperlukan tindakan agar tidak ada corealsi antar peubah (multicolinearity), maka harus diuji VIF, agar bisa baik yaah sehingga Profit bisa meningkat yaa

```
angka_dat<- cor(dataku[,num_kol])
angka_dat_tr<- cor(dataku_transform[,num_kol])


corrplot(angka_dat,main="untranform")
```




```
corrplot(angka_dat_tr,main="tranform")
```



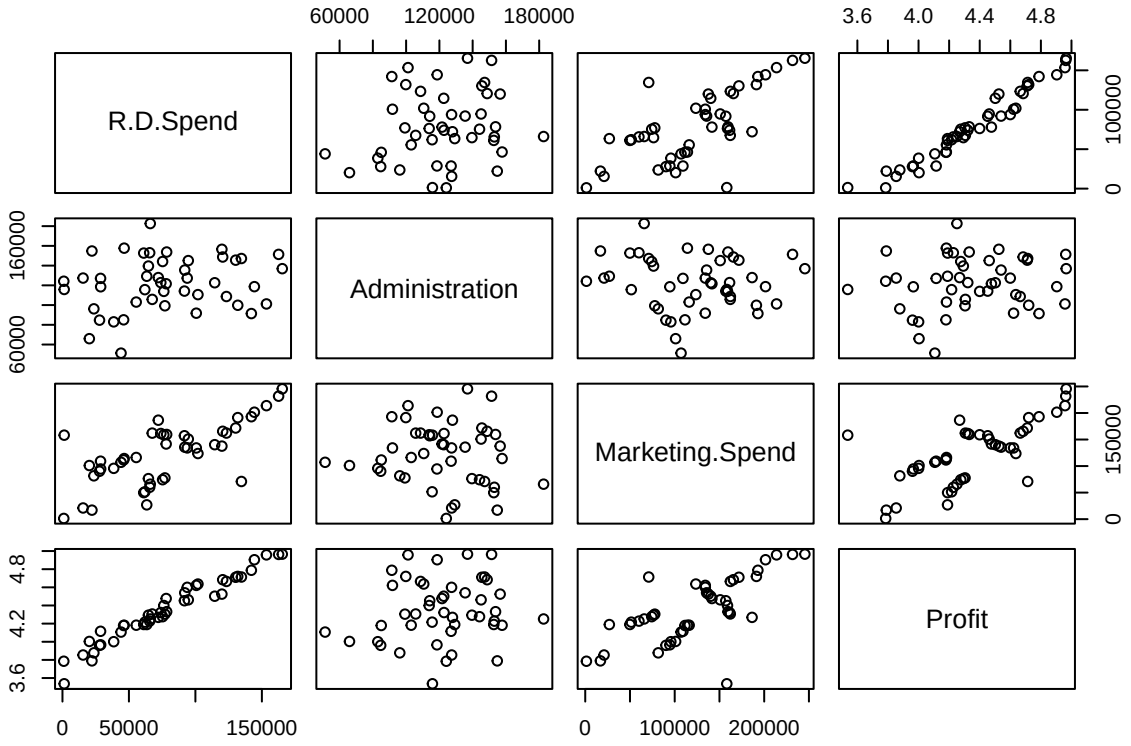
Kurang ada perbedaan yang mencolok ya

Tambahan

```
dataku_transform$Marketing.Spend <- transformTukey(dataku_transform$Marketing.Spend,
                                                   plotit = FALSE)
```

```
##
##     lambda      W Shapiro.p.value
## 439   0.95 0.9881           0.916
##
## if (lambda >  0){TRANS = x ^ lambda}
## if (lambda == 0){TRANS = log(x)}
## if (lambda <  0){TRANS = -1 * x ^ lambda}
```

```
pairs(dataku_transform[,num_kol])
```



Disini Ada suatu Perbedaan yaa