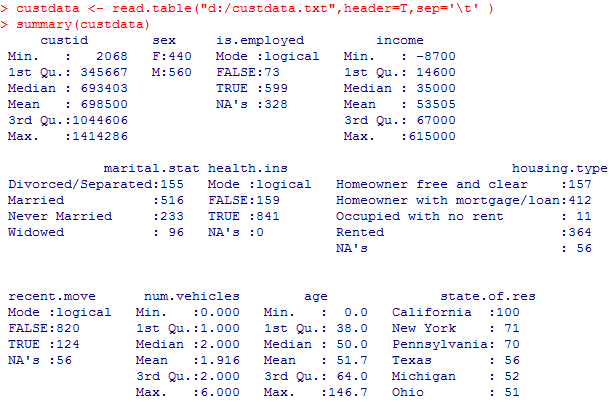
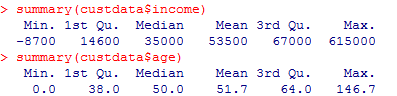
Data Exploration

Disini data yang digunakan adalah data Asuransi yaitu Custdata Ada beberapa langkah yang dilakukan untuk Eksplorasi Data yaitu

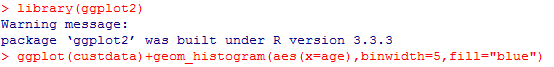
1. Loading data yaitu dengan cara yaitu

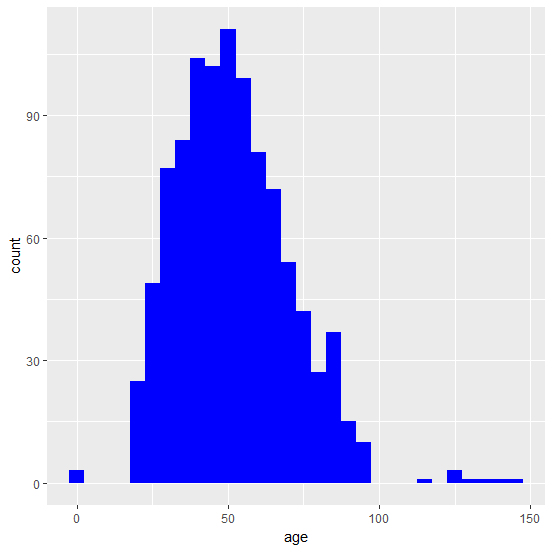


1. Eksplorasi CustData



Dimana Custdata Income yang maksimal 615000 serta Custdata Age yang maksimal 146.7. kemudian dengan menggunakan library (ggplot2) maka disini dapat diperoleh informasi soal eksplorasi data grafik yang mencerminkan distribusi data sebagai berikut.

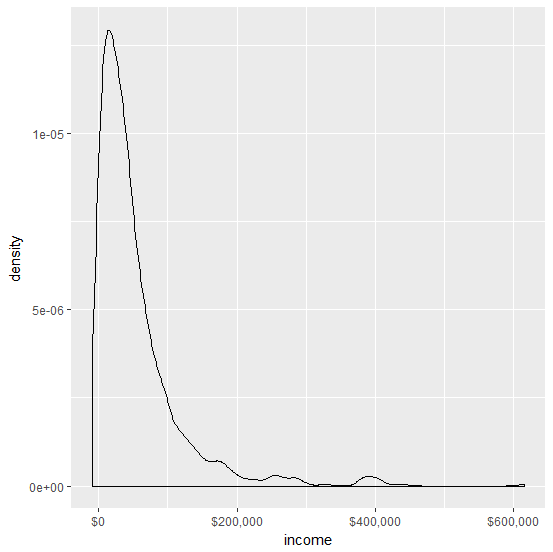




Dimana dari grafik tersebut maka diketahui jumlah pemakai asuransi tertinggi berada pada umur sekitar 50.

Setelah itu, kita melanjuti pada distribusi density income sebagai berikut dengan code sebagai berikut

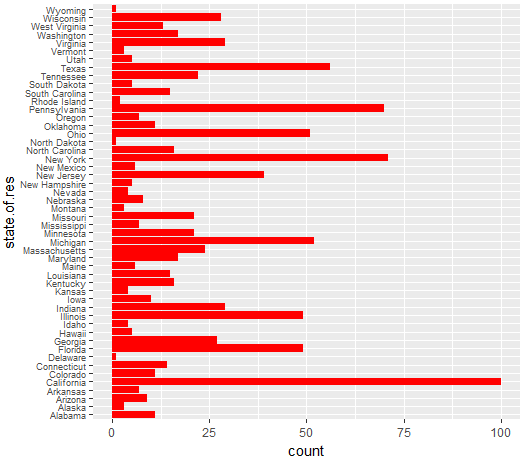
ggplot(custdata) + geom\_density(aes(x=income)) + scale\_x\_continuous(labels=dollar)



Densitas tertinggi berada pada $0 sampai $100000 untuk income pengguna asuransi.

Setelah itu untuk distribusi wilayah pengguna asuransi dapat ditunjukan dengan gambar dan code sebagai berikut

ggplot(custdata) + geom\_bar(aes(x=state.of.res), fill="red") + coord\_flip() + theme(axis.text.y=element\_text(size=rel(0.8)



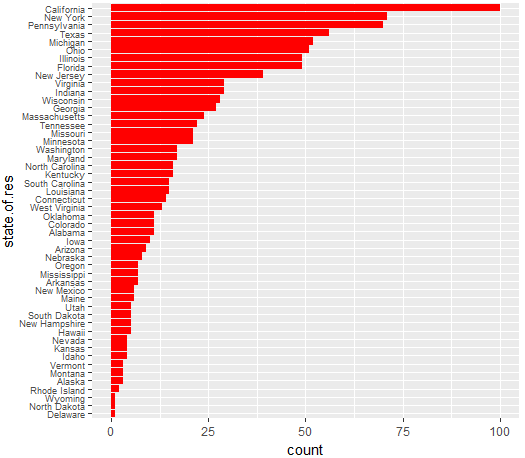
Dimana grafik ini merupakan jumlah pengguna asuransi berdasarkan daerah setelah itu agar mengetahui distribusi data lebih dalam maka data perlu di sort dengan code sebagai berikut.

statesums <- table(custdata$state.of.res)

statef <- as.data.frame(statesums)

colnames(statef)<-c("state.of.res", "count")

statef <- transform(statef, state.of.res=reorder(state.of.res, count))



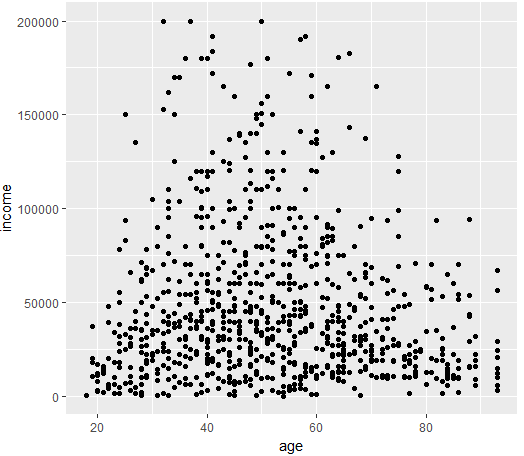
Dimana pengguna tertinggi berada di California dan terendah berada di Delaware.

Selanjutnya dengan mengambil sampel data untuk age 0<age<100 dan income>0 dengan code dan gambar sebagai berikut.

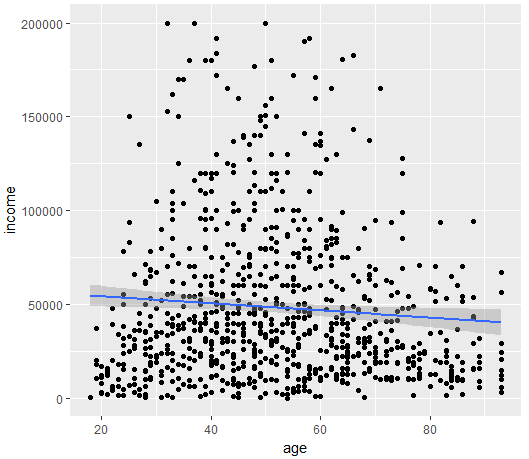
custdata2 <- subset(custdata,(custdata$age > 0 & custdata$age < 100 & custdata$income > 0))

cor(custdata2$age, custdata2$income)

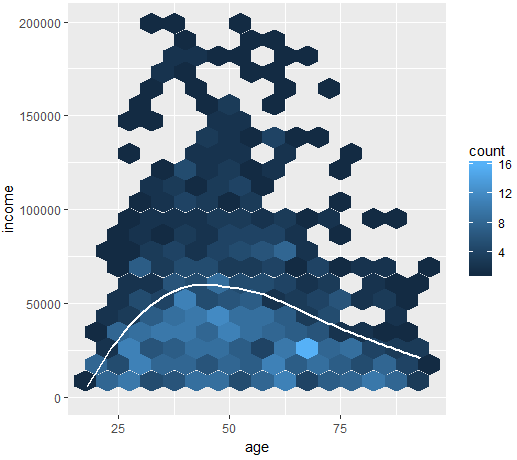
ggplot(custdata2, aes(x=age, y=income)) + geom\_point() + ylim(0, 200000)



Dimana gambar ini merupakan distribusi sebaran data antara age dan income. Selanjutnya dengan menggunakan regresi linear maka akan diperoleh grafik sebagai berikut.

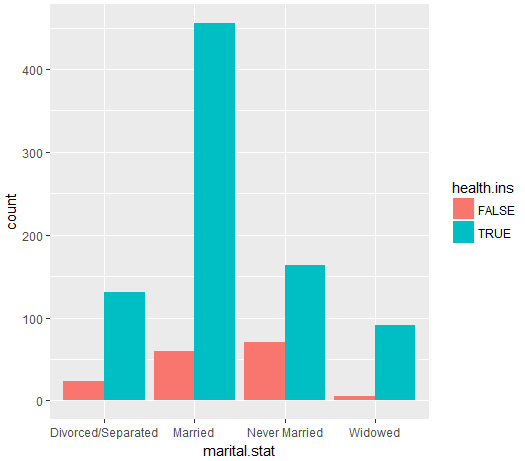


Namun disini dengan menggunakan linear maka disini model regresi tidak begitu akurat. Selanjutnya dengan trend smoothing dan kontur hexbin maka ditunjukan gambar sebagai berikut.



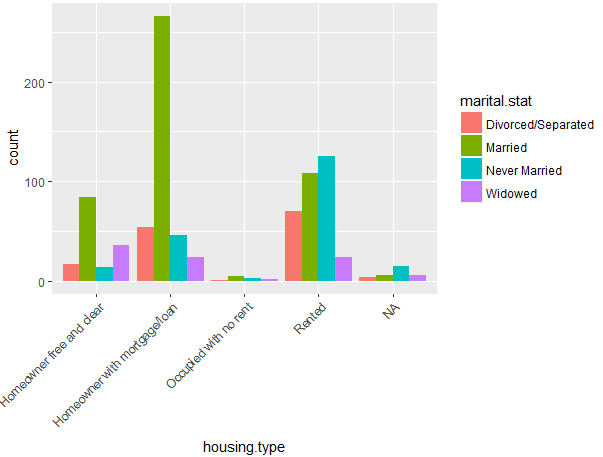
Disini kita dapat melihat kontur frekuensi data diantara age dan income dengan tren smoothing.

Selanjutnya eksplorasi data dengan status marital dapat ditunjukan gambar sebagai berikut.

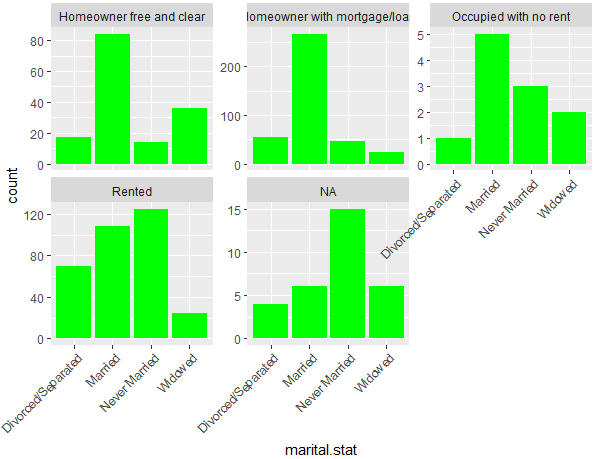


Dari gambar tersebut diperoleh informasi bahwa pengguna asuransi tertinggi ada berada pada yang telah menikah sedangkan pada yang bukan pengguna tertinggi pada yang belum menikah dan yang terendah baik itu pengguna maupun bukan pengguna asuransi berada pada duda maupun janda.

Selanjutnya berdasarkan status rumahnya maka dapat diperoleh grafik sebagai berikut.



Disini dapat disimpulkan bahwa pengguna asuransi tertinggi berada pada pemilik rumah dengan pinjaman yang sudah menikah dan terendah pada pemilik rumah bekerja yang bukan sewa.



Dan grafik ini menunjukan bagian rinci dari housing type dan marital status. Berdasarkan grafik tersebut yang teringgi berada pada status yang telah menikah untuk pemilik rumah yang bebas pinjaman, pemakai pinjaman dan bekerja dengan tanpa menyewa. Selanjutnya yang paling terendah pada pemilik bukan sewa yang sudah bercerai, pemilik yang menyewa sudah duda atapun janda dan sebagainya.