Density Estimation and visualization

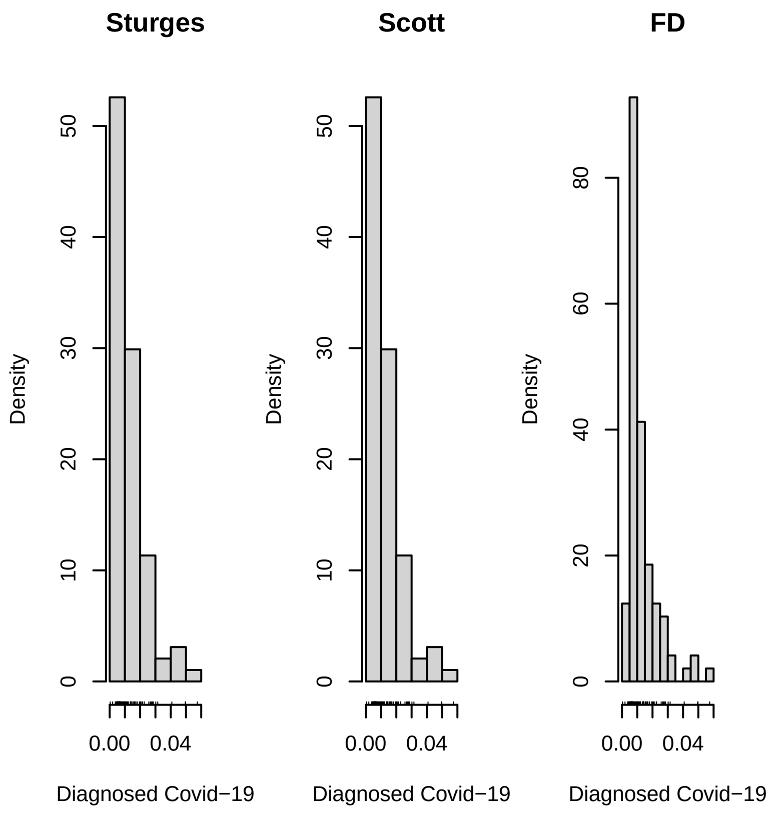
Dennis Do

10/12/21

I. Single dimension

1)

```
load("C:\\Users\\Dennis\\Downloads\\covid_sd_20201001.RData")
methods <- c("Sturges","Scott", "FD")
layout(matrix(1:3, 1,3))
for (method in methods){
  hist(as.numeric(covid$case_count_proportion), breaks = method, freq = F, xlab = "Diagnosed Covid-19",
  rug(covid$case_count_proportion)
}
```



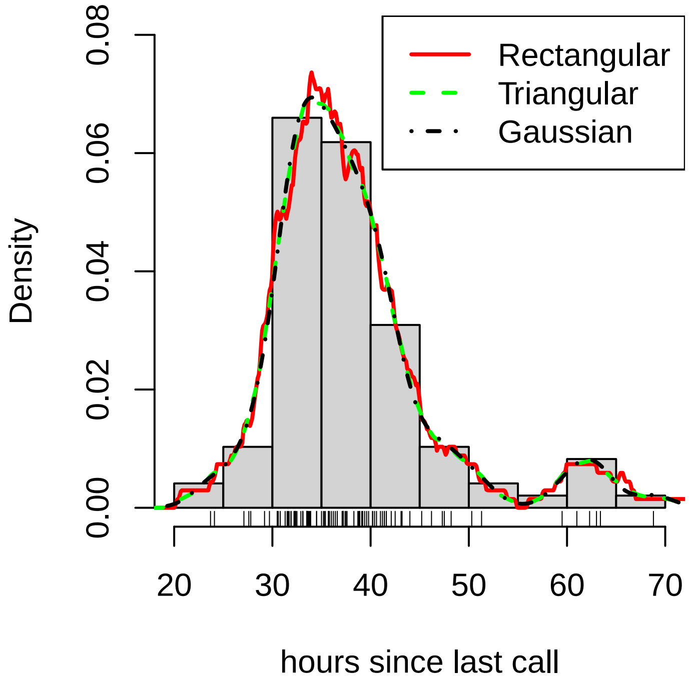
2) The default kernel used for the r density() function is gaussian.

3)

```
density(covid$median_age)
```

```
##
## Call:
##  density.default(x = covid$median_age)
##
## Data: covid$median_age (97 obs.);    Bandwidth 'bw' = 2.018
##
##        x                y
##  Min.   :17.65   Min.   :2.304e-05
##  1st Qu.:31.95   1st Qu.:2.134e-03
##  Median :46.25   Median :6.778e-03
##  Mean   :46.25   Mean   :1.746e-02
##  3rd Qu.:60.55   3rd Qu.:2.534e-02
##  Max.   :74.85   Max.   :6.941e-02
```

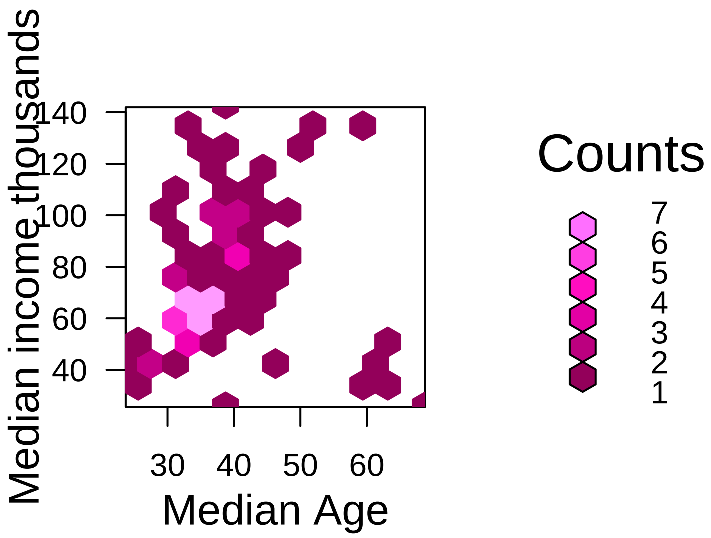
```
hist_kde <- hist(covid$median_age, breaks = 12, freq = F, xlab = "hours since last call", main = "", yl

rug(covid$median_age)
density_calls <- density(covid$median_age, kernel = "rectangular")
lines(density_calls, lwd = 2, col = "red")
density_calls2 <- density(covid$median_age, kernel = c("triangular"))
lines(density_calls2, lwd = 2, col = "green", lty = 2)
density_calls3 <- density(covid$median_age, kernel = c("gaussian"))
lines(density_calls3, lwd = 2 , col = "black", lty = 4)
legend("topright", lwd = 2, lty = c(1,2,4), col = c("red", "green", "black"), legend = c("Rectangular",
```



Two Dimensions

4)

```r
library(hexbin)
adj <- covid$median_income / 1000
hex <- hexbin(covid$median_age, adj , shape =1 , xbin = 12, xlab = "Median Age", ylab = "Median income 
plot(hex, colramp = function(n)magent(n, beg=50, end=200))
```

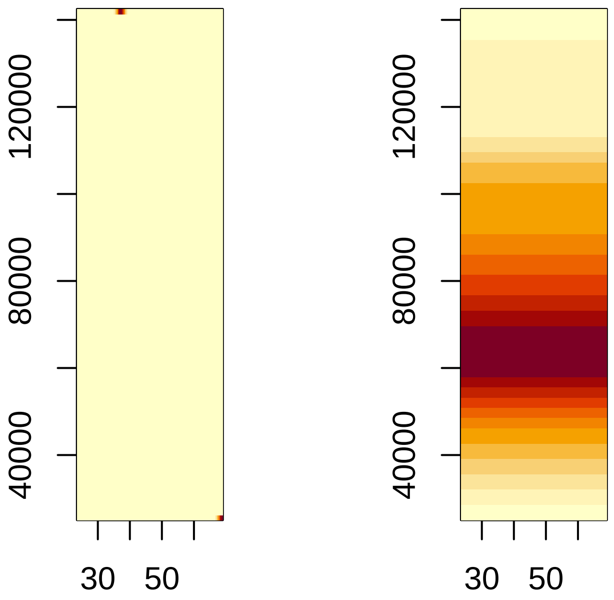


5) colramp = function(n)magent(n, beg=50, end=200)

6) The default method is gaussian kernel density estimate of the bivariate geyser. The bandwith with normal reference would be used for median age and median income.

7)

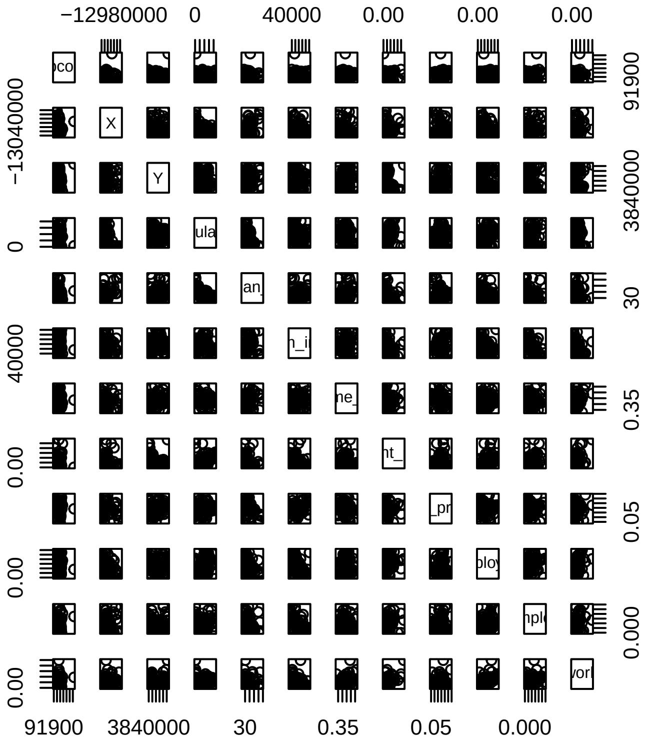
```
library(MASS)
par(mfrow=c(1,2))
image(kde2d(covid$median_age,covid$median_income,4,100))
image(kde2d(covid$median_age,covid$median_income,30000,100))
```



Several dimensions

8)

```
data(iris)
pairs(covid)
```



9) cor(covid) sort(cor(covid))

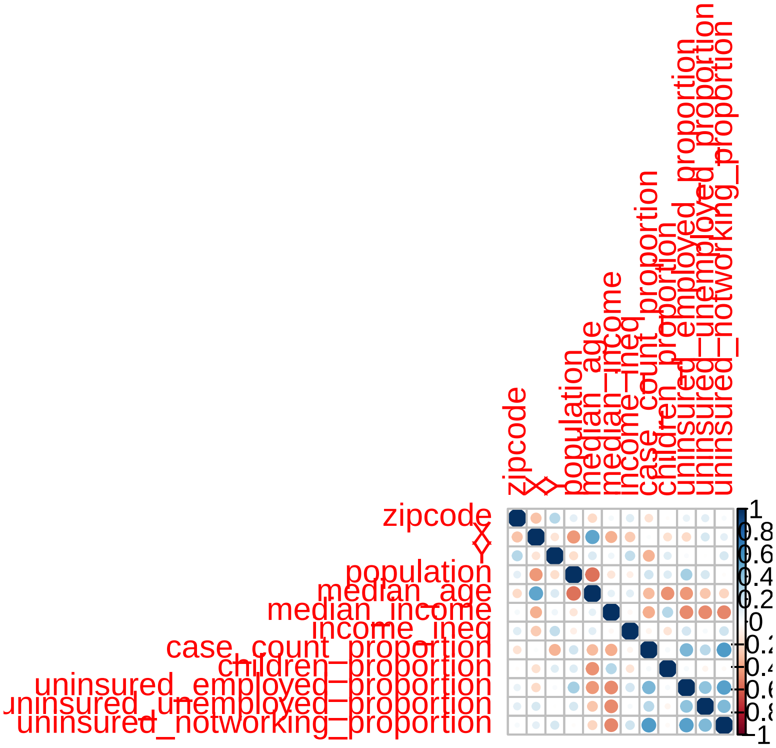
Two variables that have a non linear relationship are y and uninsured unemployed proportion. It has a value of -0.006482093 and is curved like a parabola.The relationship between the two is that it is non linear and can vary.

10)

```
library(corrplot)
```

```
## corrplot 0.90 loaded
```

```
corrplot(cor(covid))
```



Variables x and y seem to be strongly correlated with the proportion of the population with covid 19 in each zip code. The results are not surprising since zip codes is just a location compared to the variables.

11)

```
covid1 <- covid[,c("population", "median_age", "median_income", "income_ineq", "children_proportion", "
covid_scale <- scale(covid1)
p <- eigen(cov(covid_scale))$vectors
lambda <- eigen(cov(covid_scale))$values
eq_pc <- princomp(covid_scale)
sqrt(lambda)
```

```
## [1] 1.6549653 1.3336128 1.0065230 0.9304183 0.7557678 0.6201626 0.5892770
## [8] 0.5484213
```

```
eq_pc$sdev
```

```
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7    Comp.8
## 1.6464125 1.3267207 1.0013213 0.9256099 0.7518620 0.6169576 0.5862316 0.5455871
```

```
p
```

```
##              [,1]        [,2]         [,3]         [,4]         [,5]         [,6]
## [1,]  0.28057510  0.4070543  0.16377872  0.650769320 -0.03881255  0.06178961
## [2,] -0.36026141 -0.5044177  0.06991737 -0.008427079  0.01145973 -0.14865659
## [3,] -0.41261780  0.3151184 -0.33728187 -0.009304123 -0.21099141  0.70352296
## [4,]  0.07987337 -0.3099919 -0.78379152  0.376227376 -0.27110667 -0.20148805
## [5,]  0.03217692  0.5601312 -0.34602157 -0.475654291 -0.05084127 -0.47944932
## [6,]  0.49903988 -0.0123948 -0.17484714  0.104032028  0.37296768  0.05218862
## [7,]  0.42396087 -0.1207688  0.21477413 -0.202420899 -0.81885217  0.07655577
## [8,]  0.43015944 -0.2366205 -0.20921163 -0.395917648  0.26103153  0.44730883
##              [,7]        [,8]
## [1,]  0.269184125  0.47694418
## [2,] -0.172593628  0.74752683
## [3,] -0.244403736  0.13193263
## [4,]  0.133874186 -0.09813538
## [5,]  0.003668236  0.32675644
## [6,] -0.748970243  0.08140280
## [7,] -0.186957495  0.08507151
## [8,]  0.473468982  0.25701704
```

```
print(eq_pc$loadings, cutoff = 0)
```

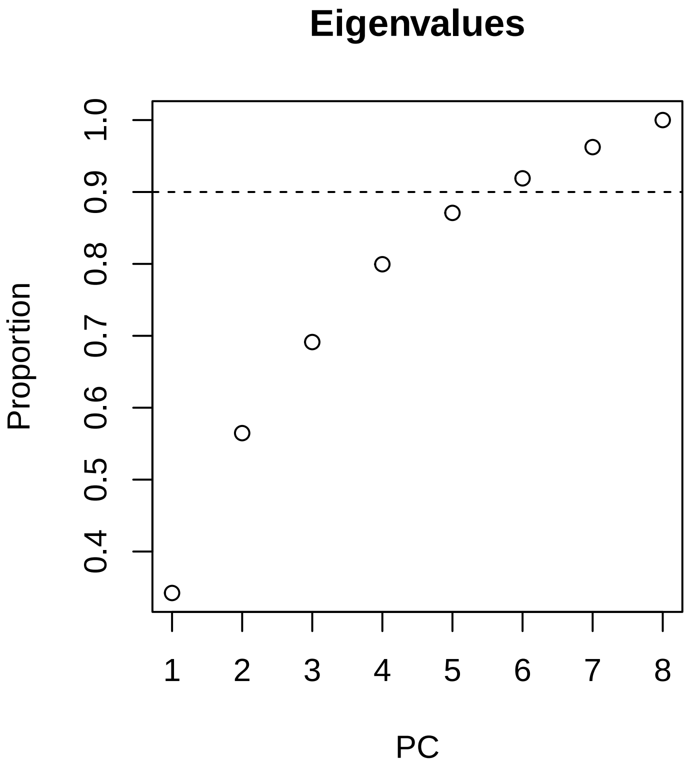
```
##
## Loadings:
##                                 Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## population                       0.281  0.407  0.164  0.651  0.039  0.062
## median_age                      -0.360 -0.504  0.070 -0.008 -0.011 -0.149
## median_income                   -0.413  0.315 -0.337 -0.009  0.211  0.704
## income_ineq                      0.080 -0.310 -0.784  0.376  0.271 -0.201
## children_proportion              0.032  0.560 -0.346 -0.476  0.051 -0.479
## uninsured_employed_proportion    0.499 -0.012 -0.175  0.104 -0.373  0.052
## uninsured_unemployed_proportion  0.424 -0.121  0.215 -0.202  0.819  0.077
## uninsured_notworking_proportion  0.430 -0.237 -0.209 -0.396 -0.261  0.447
##                                 Comp.7 Comp.8
## population                       0.269  0.477
## median_age                      -0.173  0.748
## median_income                   -0.244  0.132
## income_ineq                      0.134 -0.098
## children_proportion              0.004  0.327
## uninsured_employed_proportion   -0.749  0.081
## uninsured_unemployed_proportion -0.187  0.085
## uninsured_notworking_proportion  0.473  0.257
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.125  0.125  0.125  0.125  0.125  0.125  0.125  0.125
## Cumulative Var  0.125  0.250  0.375  0.500  0.625  0.750  0.875  1.000
```

The dominant variables in the first principle component is uninsured_employed_proportion, uninsured_unemployed_proportion and uninsured_notworking_proportion.Zip codes relation is that they have the most impact for pc1.

12) The dominant variables in Comp 2 are children population and population.

13)

```
eq_eigen_all <- eigen(cov(covid_scale))
plot(cumsum(eq_eigen_all$values)/sum(eq_eigen_all$values),
xlab = "PC", ylab = "Proportion", main = "Eigenvalues")
abline(h = 0.90, lty = 2)
```



At least 5 PC is needed to capture at least 90% of the variation.

14) This suggests that the relation between covid 19 cases and the explanatory variables are that the components affect the proportion of the population with covid the most like in the first PC.