PSTAT 105 HW 5 Question 3

Matthew Xu (5752811)

19 Feburary 2021

```
# libraries
library(tidyverse)

# scan in Carbajal data
Carbajal <- scan("Carbajal.txt", skip = 1)
```

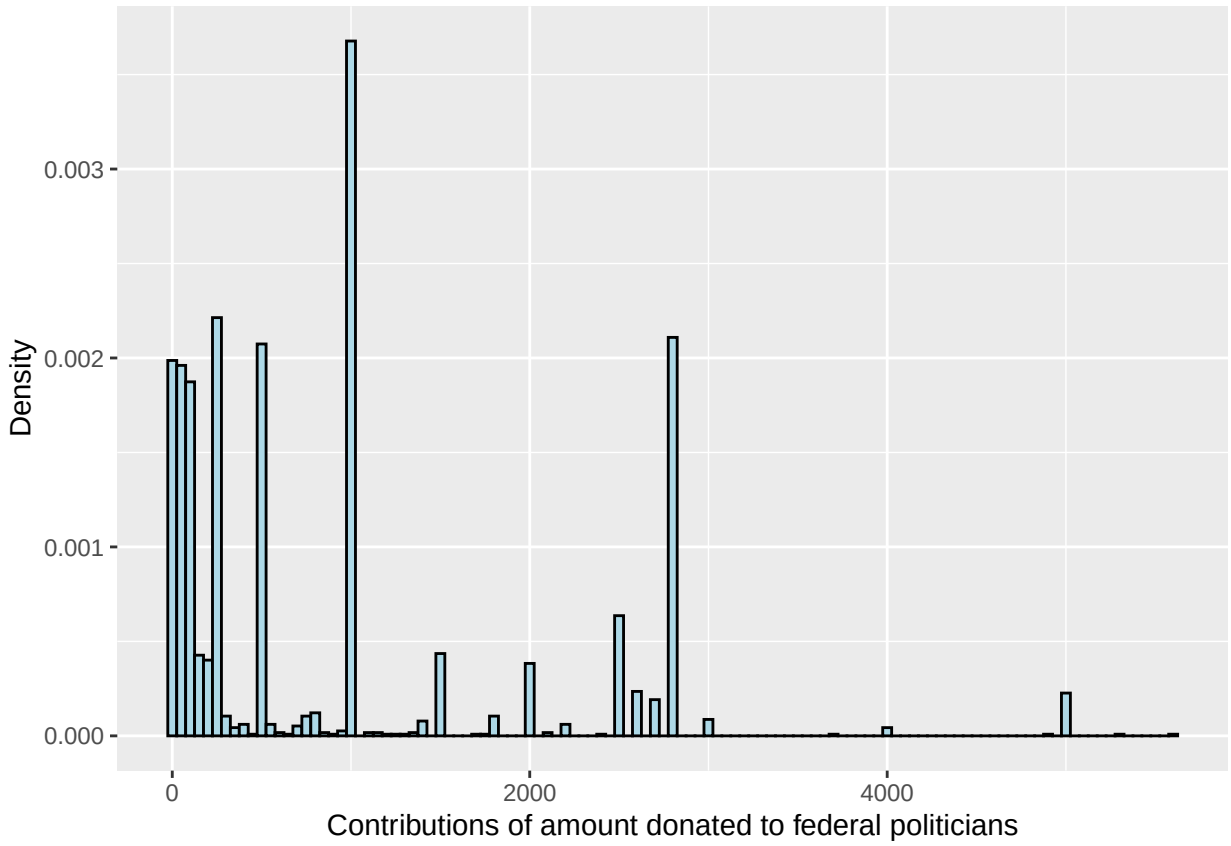
Please submit your answers to this questions as a typed report including input and output from your R code.

3. The Federal Election Commission makes available the amount of money donated to federal politicians. The data file Carbajal.txt contains the dollar amounts of 2,295 contributions made to Rep. Salud Carbajal from 2019 to 2020.

(a) Create a histogram using a number of bins that you believe allows us to see the interesting details in the contribution amounts.

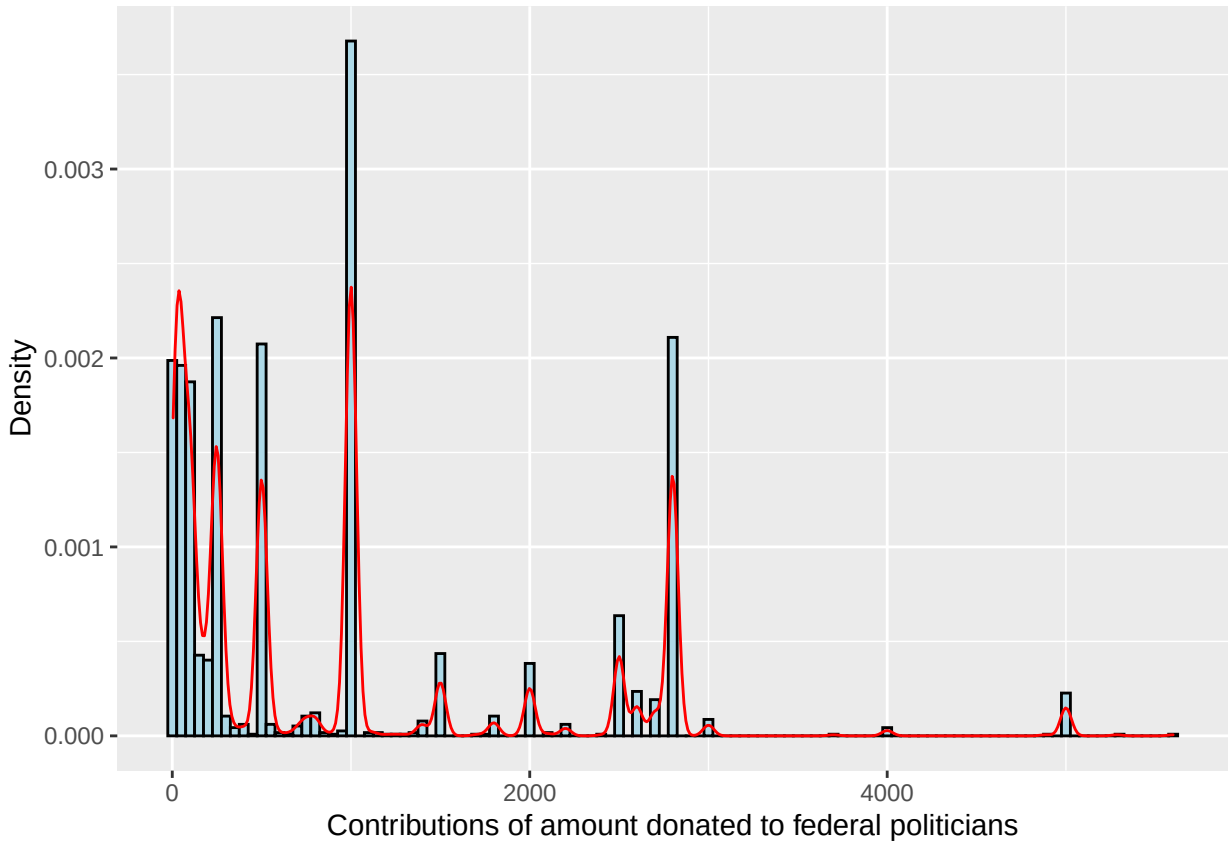
```
# histrogram with reasonable binwidth of 50
carbajal_hist <- ggplot(data.frame(Carbajal), aes(Carbajal)) + geom_histogram(binwidth = 50,
    fill = "lightblue", color = "black", aes(y = ..density..)) + labs(x = "Contributions of amo
    y = "Density")

carbajal_hist
```



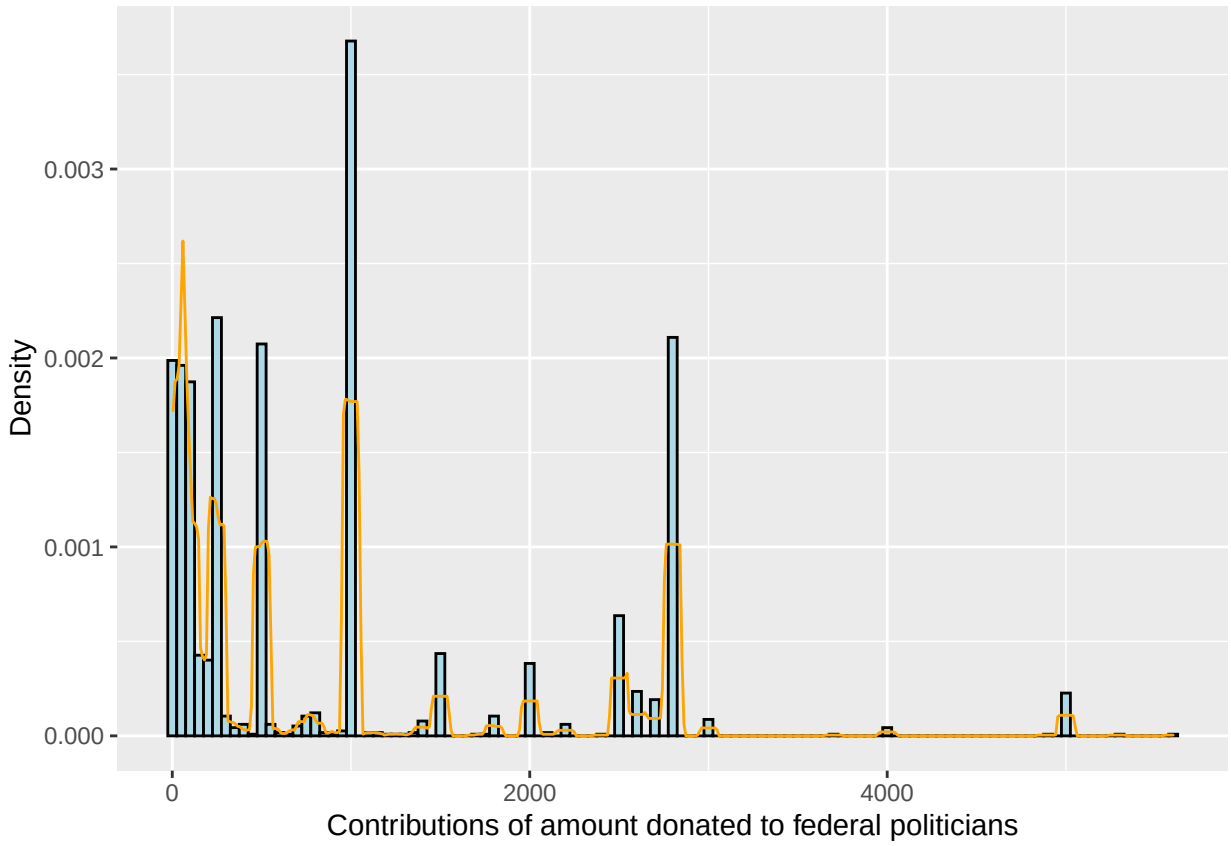
(b) Plot a copy of this histogram but include a kernel density estimate using the density or geom_density function. Set the bandwidth to 30 and use a Gaussian kernel. Keep the bandwidth the same.

```
# histrogram with gaussian kernel of banwdidth 30
carbajal_hist + geom_density(bw = 30, kernel = "gaussian", col = "red")
```

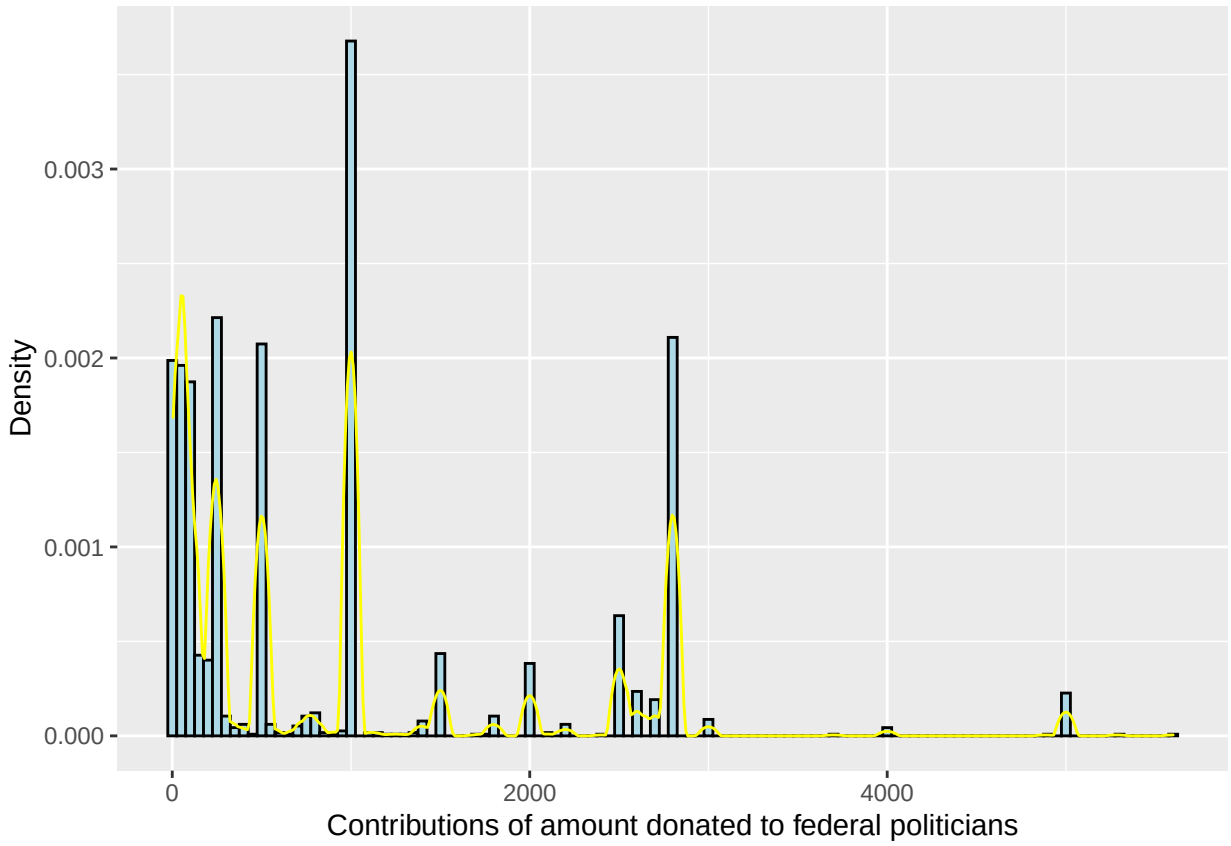


(c) Repeat this graph but use the Rectangular and then the Epanechnikov kernels.

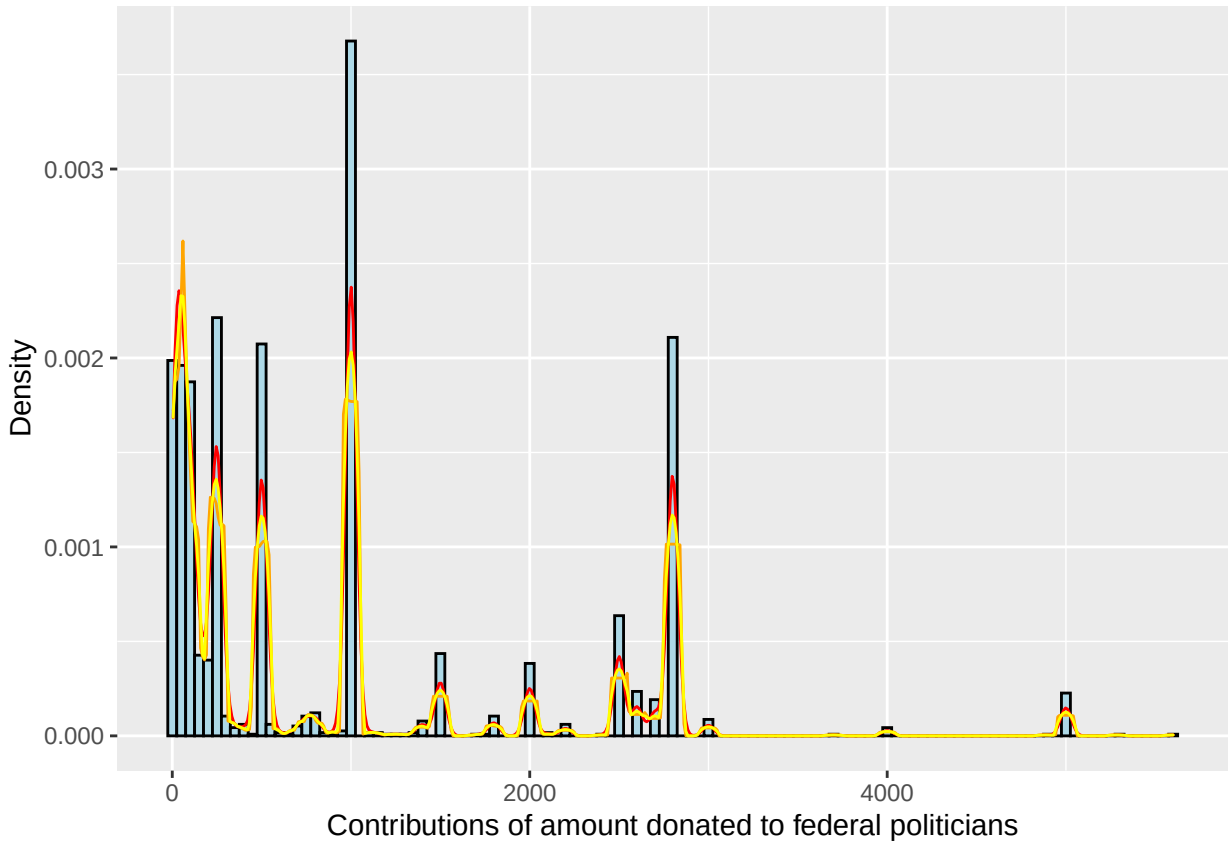
```
# histrogram with rectangular kernel of banwdidth 30
carbajal_hist + geom_density(bw = 30, kernel = "rectangular", col = "orange")
```




```
# histrogram with Epanechnikov kernel of banwdidth 30
carbajal_hist + geom_density(bw = 30, kernel = "epanechnikov", col = "yellow")
```



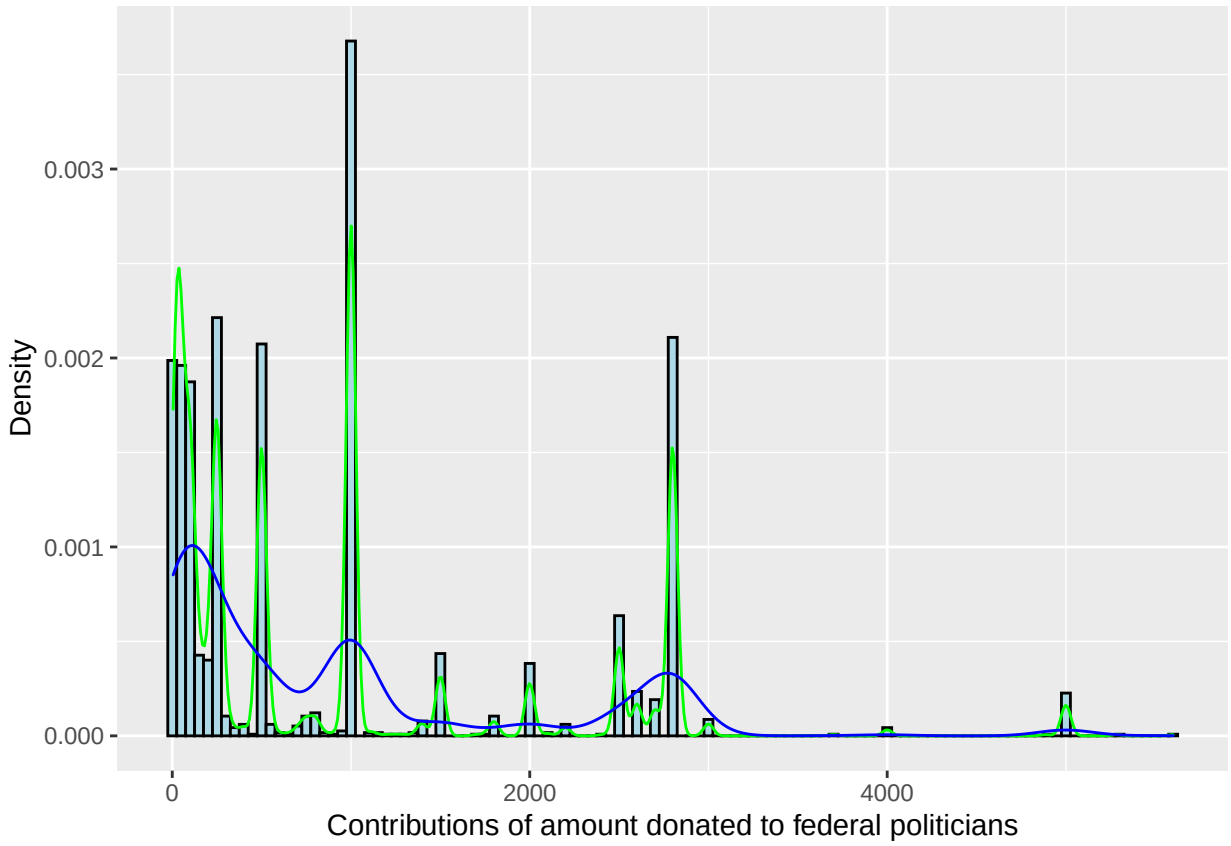

```
# together
carbajal_hist + geom_density(bw = 30, kernel = "gaussian", col = "red") + geom_density(bw = 30
    kernel = "rectangular", col = "orange") + geom_density(bw = 30, kernel = "epanechnikov",
    col = "yellow")
```



(d) What difference does the kernel shape make? The epanechnikov and gaussian curves appear to be smoother, as they have less points of contributions that appear to flatten at the peak of the curves. However, the rectnagular kernel appears to match the histogram at each peak the closest and the best, even though it appears to be less smooth. The flexibility appears to be best at the epanechnikov kernel still, as it is the smoothest, with density curves at peaks that are not too far off. The rectangular kernel may be overfitting, as it matches more closely, making epanechnikov probabily the most flexible of the three kernels.

(e) Produce another plot of the data, but now use automated bandwidth options bw="ucv" and bw="nrd" to draw two density estimates on the histogram.

```
# histrogrm with ucv and nrd
carbajal_hist + geom_density(bw = "ucv", kernel = "gaussian", col = "green") + geom_density(bw
    kernel = "gaussian", col = "blue")
```



(f) Find a bandwidth for the density estimate that you think works best and estimate the density at 50 and 1500.

```
# bandwidth estimators least squares cross validation
bw.ucv(x = Carbajal)
```

```
## [1] 26.90453
```

```
# normal scale bandwith selector
bw.nrd(x = Carbajal)
```

```
## [1] 151.4562
```

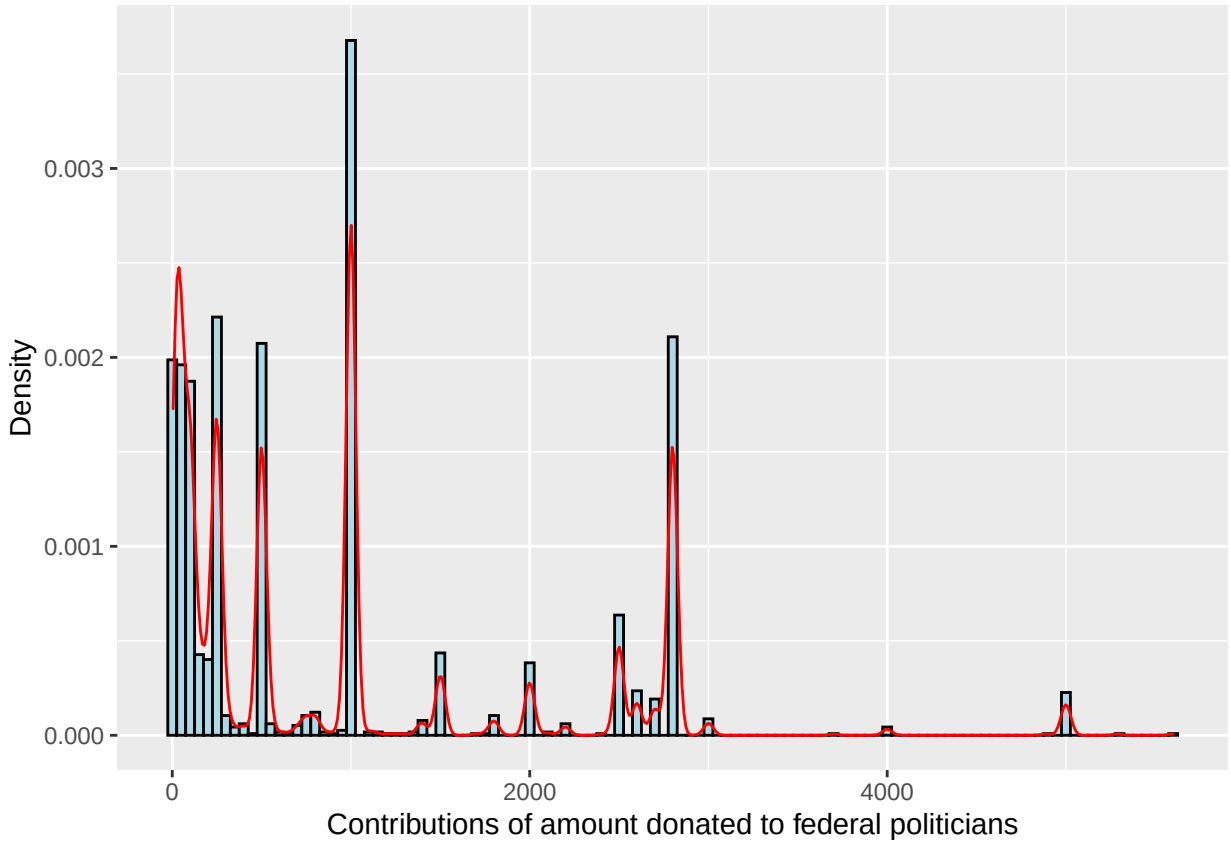
```
# direct plug in selector
bw.SJ(x = Carbajal, method = "dpi")
```

```
## [1] 34.77741
```

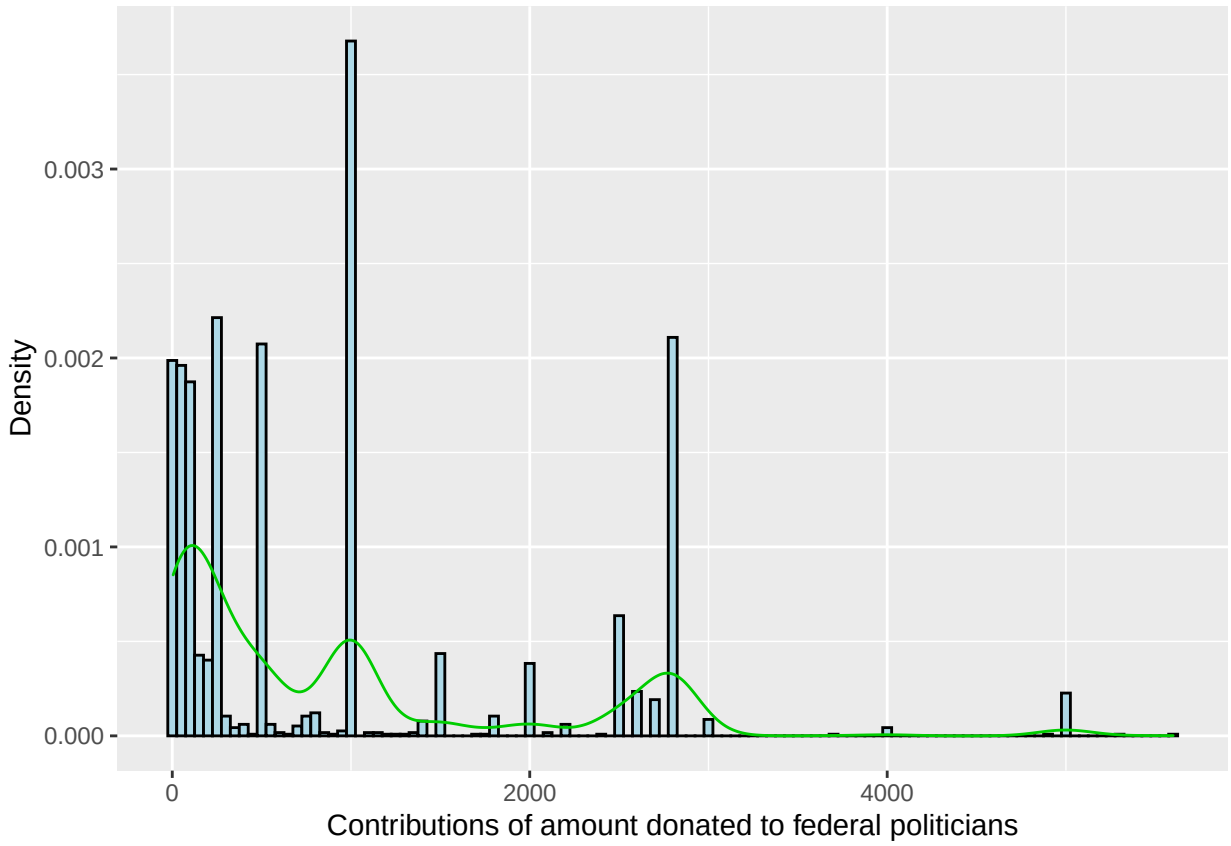
```
# biased cross validation selector
bw.bcv(x = Carbajal)
```

```
## [1] 257.0613
```

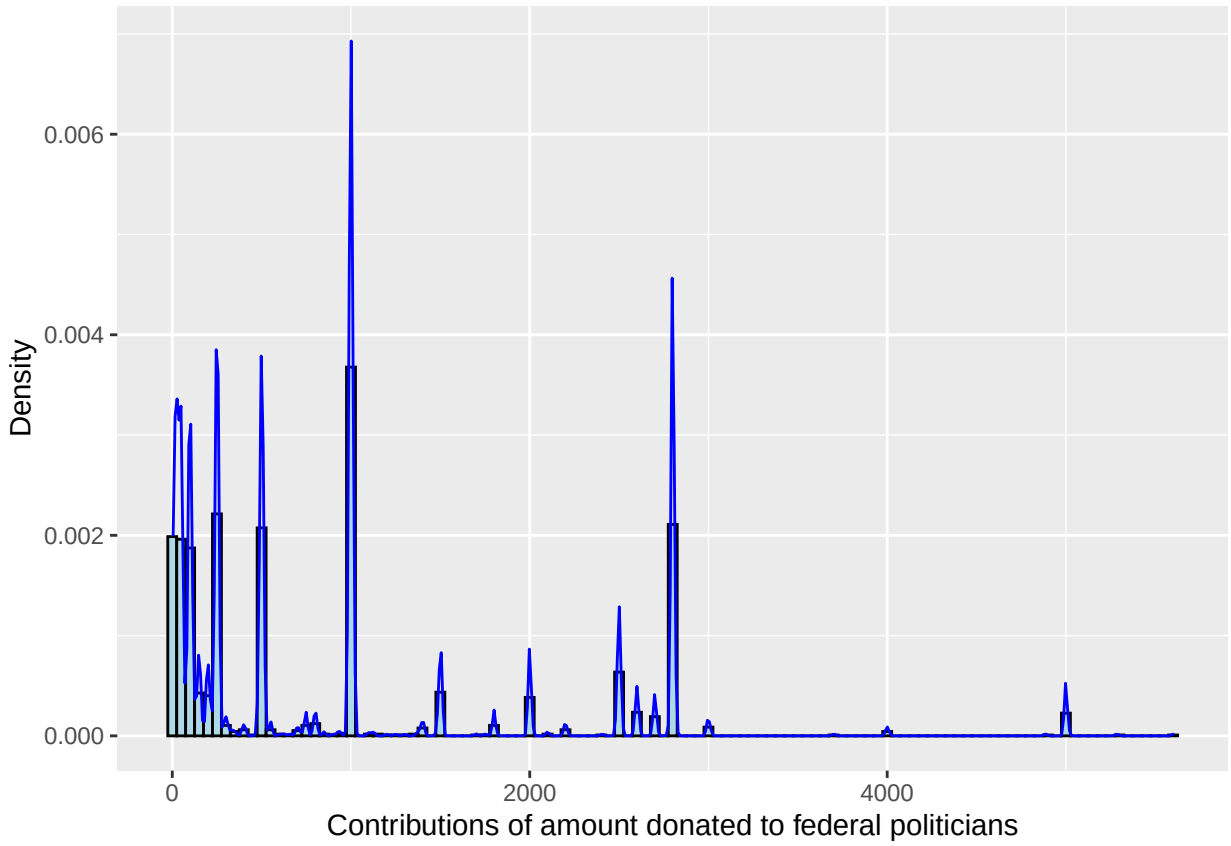
```
carbajal_hist + geom_density(bw = "ucv", kernel = "gaussian", col = 2)
```



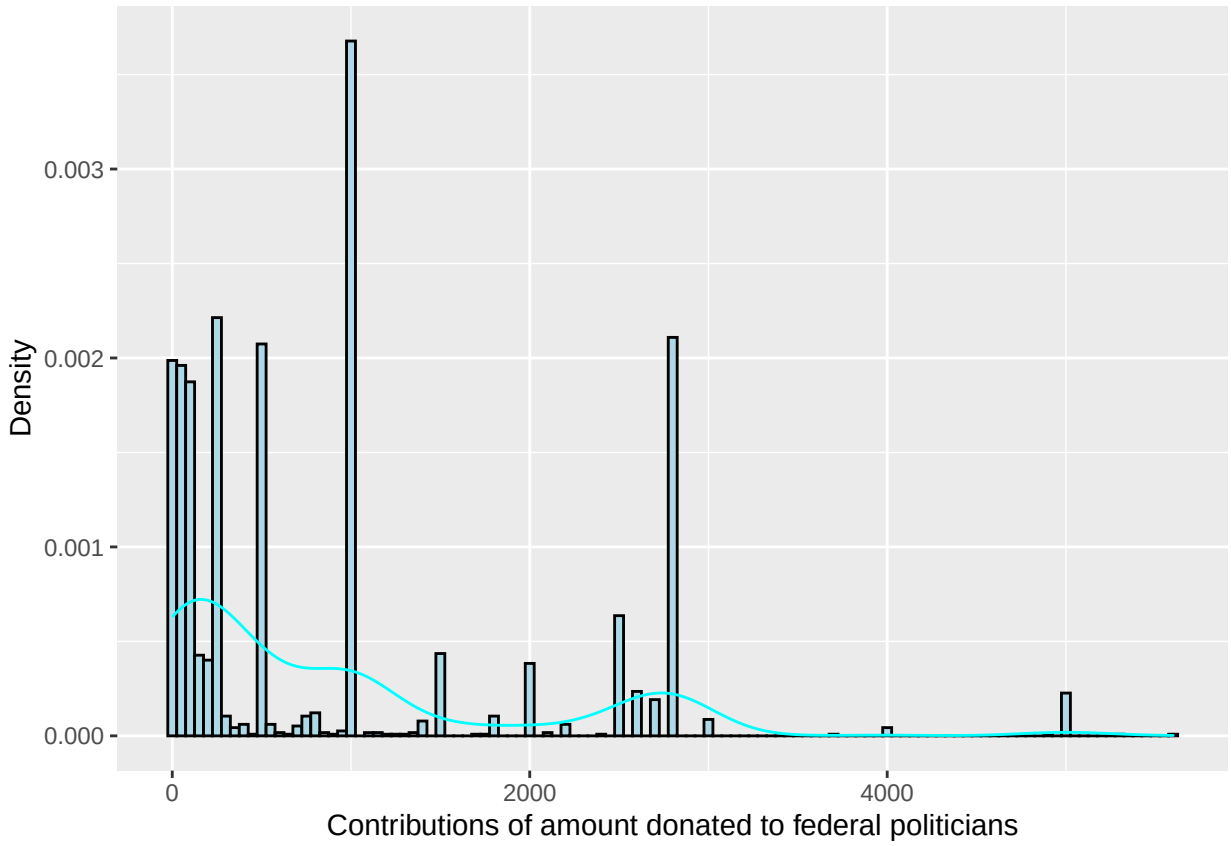

```
carbajal_hist + geom_density(bw = "nrd", kernel = "gaussian", col = 3)
```



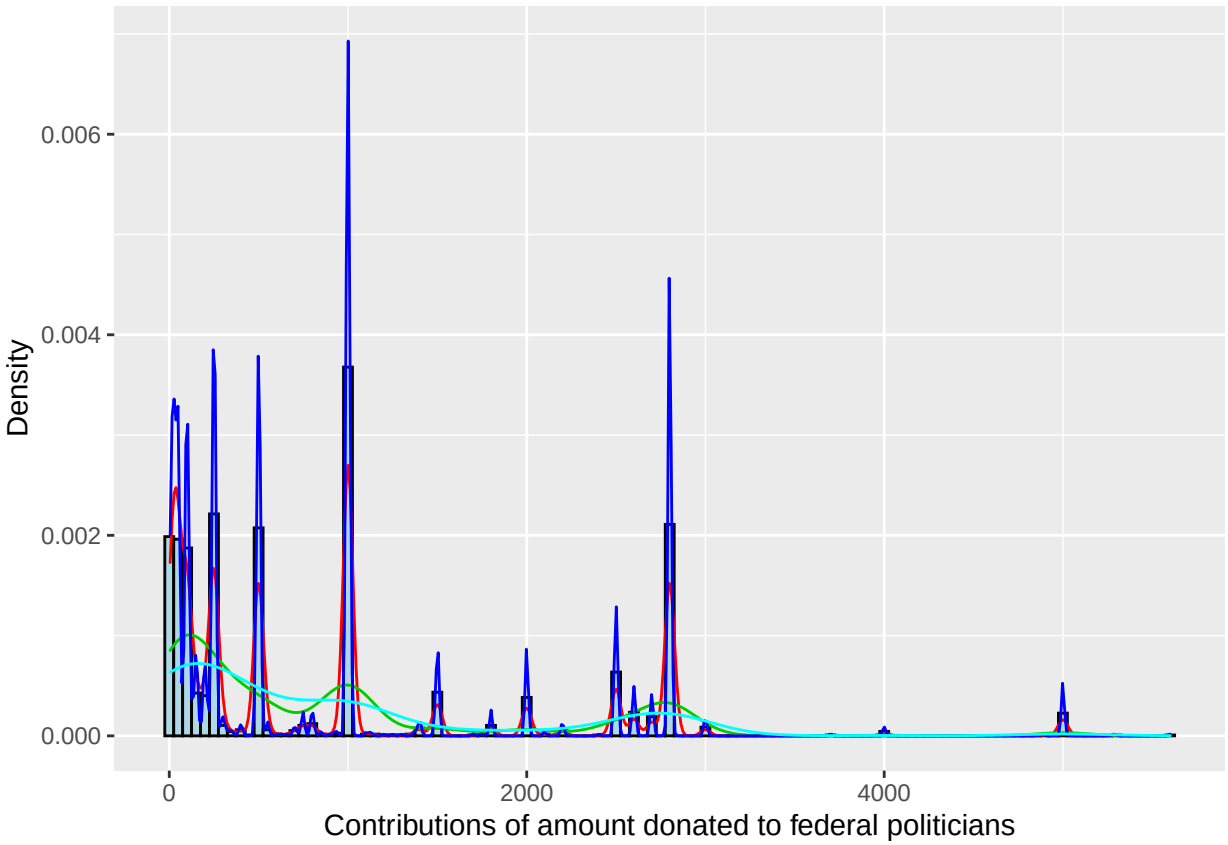

```
carbajal_hist + geom_density(bw = "SJ", kernel = "gaussian", col = 4)
```




```
carbajal_hist + geom_density(bw = "bcv", kernel = "gaussian", col = 5)
```




```
# together
carbajal_hist + geom_density(bw = "ucv", kernel = "gaussian", col = 2) + geom_density(bw = "nr
    kernel = "gaussian", col = 3) + geom_density(bw = "SJ", kernel = "gaussian",
    col = 4) + geom_density(bw = "bcv", kernel = "gaussian", col = 5)
```



The best bandwidth estimator appears to be ucv with kernel gaussian, being the smoothest and fitting the histogram the best. At these paramaters, dendsity at 50 is about 0.002, density at 1500 appears to be about 0.0004.

(g) Does the shape of the kernel or the bandwidth have a greater effect on the resulting density estimate?

The bandwidth is usually a smoothing paramater of the kenerl chosen, scaling the kernel to the densities of the data. It does not directly generate densities and is still based on the kernel chosen itself. Therefore, in this interpretation, the kerneal has the largest effect on the produced desntiy estimates. The bandwidth merely scales these desnity values.