Exercise 8

1. Mean and median of two datasets (Theory)

Consider two datasets $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$. Note that they have different lengths. Let $\bar{x}$ be the sample mean of the first, and $\bar{y}$ the sample mean of the second. Consider the combined dataset $x_1, \ldots, x_n, y_1, \ldots, y_m$ with $m + n$ elements, obtained by concatenating the two original datasets.

a. Is it true that the sample mean of the combined dataset is equal to $\frac{\bar{x}+\bar{y}}{2}$? If yes, provide a proof, if no, provide a counterexample.

Solution. It is not the case. To see why we can keep substituting and reducing $\frac{\bar{x}+\bar{y}}{2}$. Let Z be the combined X and Y datasets, then the sample mean would be:

$$\bar{z} = \frac{(x_1 + \cdots + x_n) + (y_1 + \cdots + y_m)}{n + m}$$

Then we can see how close can we get to that with the given formula:

$$\frac{\bar{x} + \bar{y}}{2} = \frac{1}{2} \cdot \Big(\frac{(x_1 + \cdots + x_n)}{n} + \frac{(y_1 + \cdots + y_m)}{m} \Big) = \frac{1}{2} \cdot \Big(\frac{m \cdot (x_1 + \cdots + x_n) + n \cdot (y_1 + \cdots + y_m)}{n \cdot m} \Big)$$

And now it seems we can't get any further. This is not at all similar to the previous formula, and we should be able to find a counterexample easily, like the following:

$$x = (1, 2, 3, 4)$$
$$y = (7, 8, 9)$$

$$\bar{x} = \frac{(1 + 2 + 3 + 4)}{4} = 2.5$$
$$\bar{y} = \frac{(7 + 8 + 9)}{3} = 8$$

$$\bar{z} = \frac{(1 + 2 + 3 + 4) + (7 + 8 + 9)}{4 + 3} \approx 4.9$$
$$\frac{\bar{x} + \bar{y}}{2} = \frac{2.5 + 8}{2} = 5.25$$

b. Consider the case where $m = n$, i.e. the two datasets have the same size. In this special case, is the sample mean of the combined dataset equal to $\frac{\bar{x}+\bar{y}}{2}$? If yes, provide a proof, if no, provide a counterexample.

Solution: The proof is done by using the previous formula:

$$\frac{\bar{x}+\bar{y}}{2} = \frac{1}{2}\cdot\Big(\frac{m\cdot(x_1+\cdots+x_n)+n\cdot(y_1+\cdots+y_m)}{n\cdot m}\Big) \overset{\text{If n = m}}{=} \frac{1}{2}\cdot\Big(\frac{n\cdot((x_1+\cdots+x_n)+(y_1+\cdots+y_m))}{n\cdot n}\Big) =$$
$$= \frac{1}{2}\cdot\Big(\frac{(x_1+\cdots+x_n)+(y_1+\cdots+y_m)}{n}\Big) = \Big(\frac{(x_1+\cdots+x_n)+(y_1+\cdots+y_m)}{2\cdot n}\Big) =$$
$$= \Big(\frac{(x_1+\cdots+x_n)+(y_1+\cdots+y_m)}{n+m}\Big) = \bar{z}$$

c. Consider now the sample medians Med_x and Med_x of the two datasets, in the general case of $m \neq n$. Is it true that the sample median of the combined dataset is equal to $\frac{Med_x+Med_y}{2}$? If yes, provide a proof, if no, provide a counterexample.

Solution: It is not the case, consider the following datasets, and z as the combination of both:

$$\begin{aligned} x &= (1,2,3) \\ y &= (5,10,20,25,30) \\ z &= x||y = (1,2,3,5,10,20,25,30) \\ Med_x &= 2 \\ Med_y &= 20 \\ Med_z &= \frac{5+10}{2} = 7.5 \\ \frac{Med_x+Med_y}{2} &= \frac{2+20}{2} = 11 \end{aligned}$$

d. In the special case of $m = n$, is the sample median of the combined dataset is equal to $\frac{Med_x+Med_y}{2}$? If yes, provide a proof, if no, provide a counterexample.

Solution: Same as before, not the case:

$$\begin{aligned} x &= (1,2,3) \\ y &= (5,10,20) \\ z &= x||y = (1,2,3,5,10,20) \\ Med_x &= 2 \\ Med_y &= 10 \\ Med_z &= \frac{3+5}{2} = 4 \\ \frac{Med_x+Med_y}{2} &= \frac{2+10}{2} = 6 \end{aligned}$$

2. Recognizing plots (Theory)

Consider the following distributions:

- $N(0,1)$
- $N(0,8)$
- $N(5,2)$
- $Exp(2)$
- $Exp(1/2)$

The following plots report histograms, kernel density estimates, and empirical distribution functions, each for a different dataset of 150 points generated from the above distributions. For each plot, say which type of plot it is (i.e. if it's a histogram, a kernel density estimate or an empirical distribution function), and identify from which of the above distributions it was generated.

Solution: It is noticeable that not all of the possible combinations are drawn. Each dataset corresponds to the following:

Dataset 1 Empirical distribution function of $N(0,1)$

Dataset 2 Kernel density estimate of $N(0,1)$

Dataset 3 Histogram of $Exp(2)$

Dataset 4 Kernel density estimate of $Exp(1/2)$

Dataset 5 Histogram of $N(0,8)$

Dataset 6 Histogram of $N(5,2)$

Dataset 7 Kernel density estimate of $Exp(2)$

Dataset 8 Empirical distribution function of $Exp(2)$

Dataset 9 Kernel density estimate of $N(5,2)$

Dataset 10 Empirical distribution function of $N(5,2)$

Dataset 11 Empirical distribution function of $Exp(1/2)$

Dataset 12 Histogram of $Exp(1/2)$

The main points we have to focus on in order to distinguish them are:

- In a normal distribution: where is it centered and how close to the mean is all the mass.
- In an exponential: the rate of growth/decay and how 'big' are the initial values.

3. Chopsticks factory (Theory, Home Exercise)

You are running a chopstick factory: due to the production process, the chopsticks are not of the exactly same length. As part of quality control you want to check that all chopsticks have approximately the same length. Suppose you produce 2000 chopsticks each day:

a. Choose an appropriate model for the chopsticks length. *Hint*: consider the model usually used to account for random variations in productions, experimental measures, etc.

Solution: We probably want a normal distribution centered on the desired length.

b. Let x_i be the length of the i-th chopstick produced today. At the end of the day you see that $\sum_i x_i = 45999$ and $\sum_i x_i^2 = 1058019$. Estimate μ and σ^2 for the distribution you chose in point a. *Hint*: look closely at how the variance is estimated and rework the formula so to be able to use the available data.

Solution: From the Central Limit Theorem we know that the sum of many intependent identically distributed random variables tends towards a normal distribution. We can assume that the production of one chopstick does not influence the others, and that all are equally produced. Thus, we can estimate the mean and variance of our distribution as follows:

$$\mu \approx \frac{\sum_i^n x_i}{n} = \frac{45999}{2000} = 22.9995$$
$$\sigma^2 \approx \frac{\sum_i^n x_i^2 - \frac{(\sum_i^n x_i)^2}{n}}{n-1} = \frac{1058019 - \frac{45999^2}{2000}}{2000-1} \approx 0.03$$
$$\sigma \approx 0.18$$

c. Give an estimate of the probability that the length of a random chopstick is within the interval $[22.5, 23.5]$.

Solution: Using our model, we just need to calculate:

$$P(N(\mu, \sigma) \geq 23.5) - P(N(\mu, \sigma) \geq 22.5) \approx 0.995$$

4. Simple Statistics (R)

Consider the `firstchi` dataset, available in the `UsingR` package, which you can load using the `library(UsingR)` statement. Using R functions, compute the following numerical statistics for the dataset.

- the sample mean
- the sample variance
- the 30th empirical percentile
- the median
- the MAD
- the IQR

You can refer to Section 2.3 of *Using R for introductory statistics*.

Solution:

```
library(UsingR)
```

```
cat("Sample mean:", mean(firstchi), "\n")
```

```
## Sample mean: 23.97701
```

```
cat("Sample variance:", var(firstchi), "\n")
```

```
## Sample variance: 39.11574
```

```
cat("30th empirical percentile:", quantile(firstchi, 0.3), "\n")
```

```
## 30th empirical percentile: 21
```

```
cat("Median:", median(firstchi), "\n")
```

```
## Median: 23
```

```
cat("MAD:", mad(firstchi), "\n")
```

```
## MAD: 4.4478
```

```
cat("IQR:", IQR(firstchi), "\n")
```

```
## IQR: 6
```

5. Plotting distributions (R)

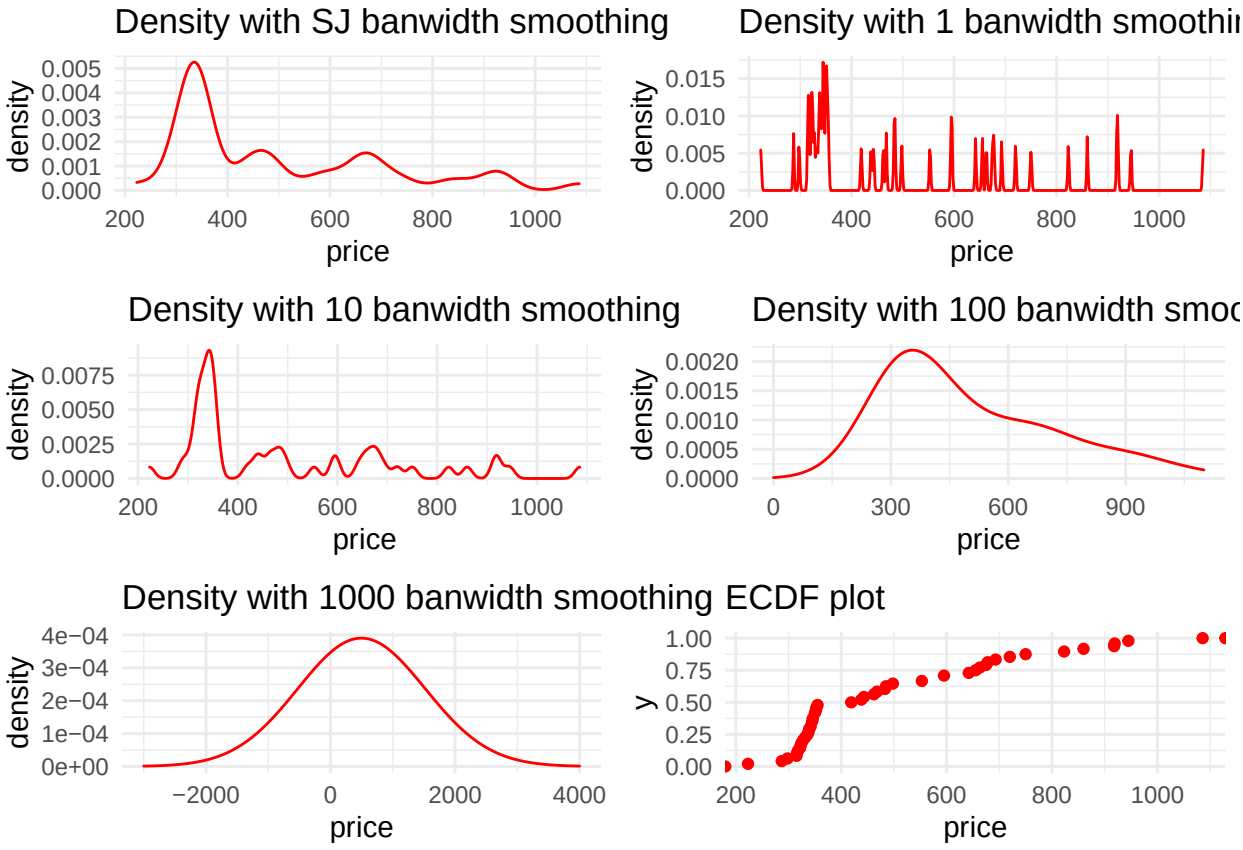
The `diamond` dataset of the `UsingR` package contains the price in Singapore dollars of 48 diamond rings, along with their size in carats.

1. Plot the kernel density estimate of prices. Try different bandwidths. How many modes are there? Look also at the empirical cumulative distribution function. Discuss your findings.

Solution:

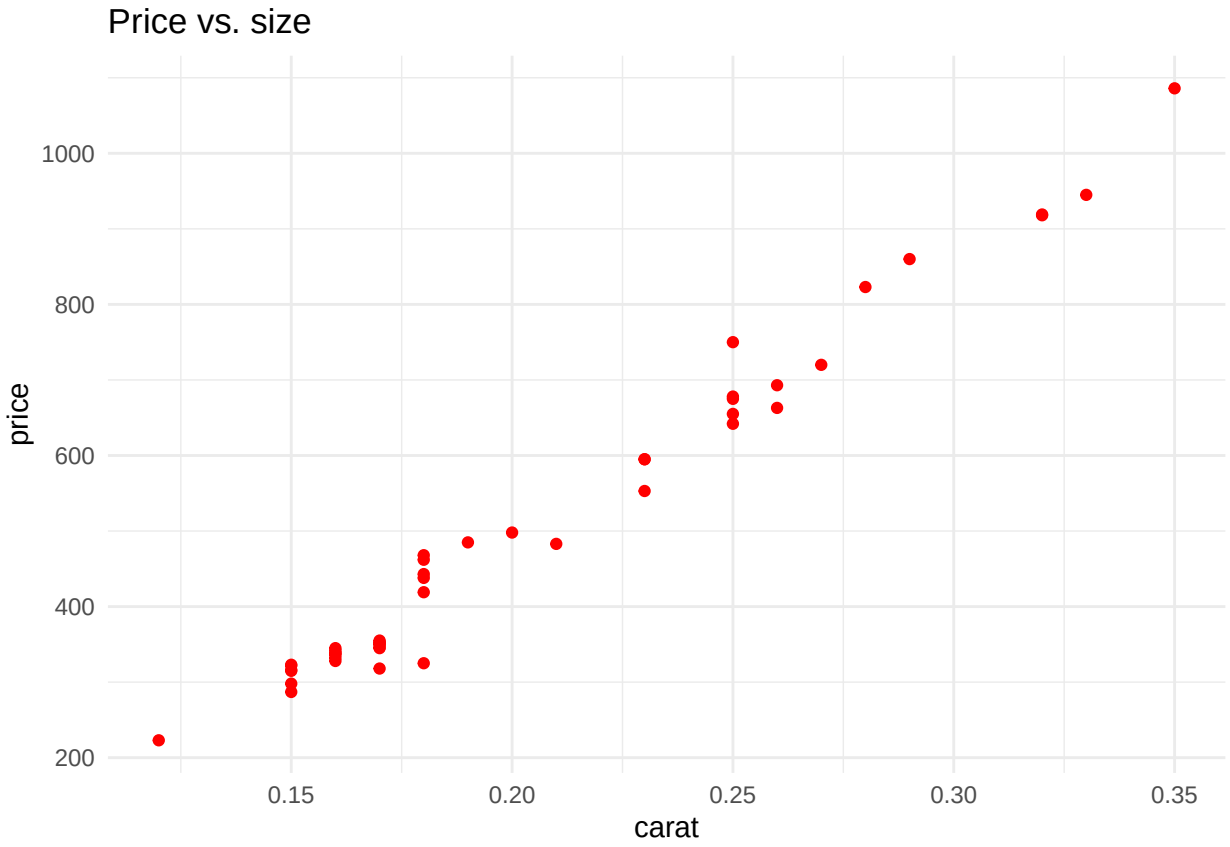
If we look at the documentation for the "density" function of R, it tells us that for the bandwith we can either use a value or one of the different strings giving a rule to choose the bandwith, those are: nrd0, nrd, ucv, bcv, SJ. By default and for historical and compatibility reasons R chooses "nrd0", although it tells us that it might not be the best option.

```r
pl1 = ggplot(diamond, aes(x = price)) +
  geom_density(bw = "SJ", color = 'red') + ggtitle("Density with SJ banwidth smoothing") +
  theme_minimal()

pl2 = ggplot(diamond, aes(x = price)) +
  geom_density(bw = 1, color = 'red') + ggtitle("Density with 1 banwidth smoothing") +
  theme_minimal()

pl3 = ggplot(diamond, aes(x = price)) +
  geom_density(bw = 10, color = 'red') + ggtitle("Density with 10 banwidth smoothing") +
  theme_minimal()

pl4 = ggplot(diamond, aes(x = price)) +
  geom_density(bw = 100, color = 'red') +
  xlim(0, 1100) + ggtitle("Density with 100 banwidth smoothing") +
  theme_minimal()

pl5 = ggplot(diamond, aes(x = price)) +
  geom_density(bw = 1000, color = 'red') +
  xlim(-3000, 4000) + ggtitle("Density with 1000 banwidth smoothing") +
  theme_minimal()

pl6 = ggplot(diamond, aes(price)) +
  stat_ecdf(geom = "point", color = 'red') + ggtitle("ECDF plot") +
  theme_minimal()

library(gridExtra)
grid.arrange(pl1, pl2, pl3, pl4, pl5, pl6)
```



2. Plot a scatterplot of prices versus sizes. Does any relation between the two quantities show up?

```
ggplot(diamond) +
  geom_point(aes(x = carat, y = price), color = 'red') + ggtitle("Price vs. size") +
  theme_minimal()
```



We can see a correlation here.

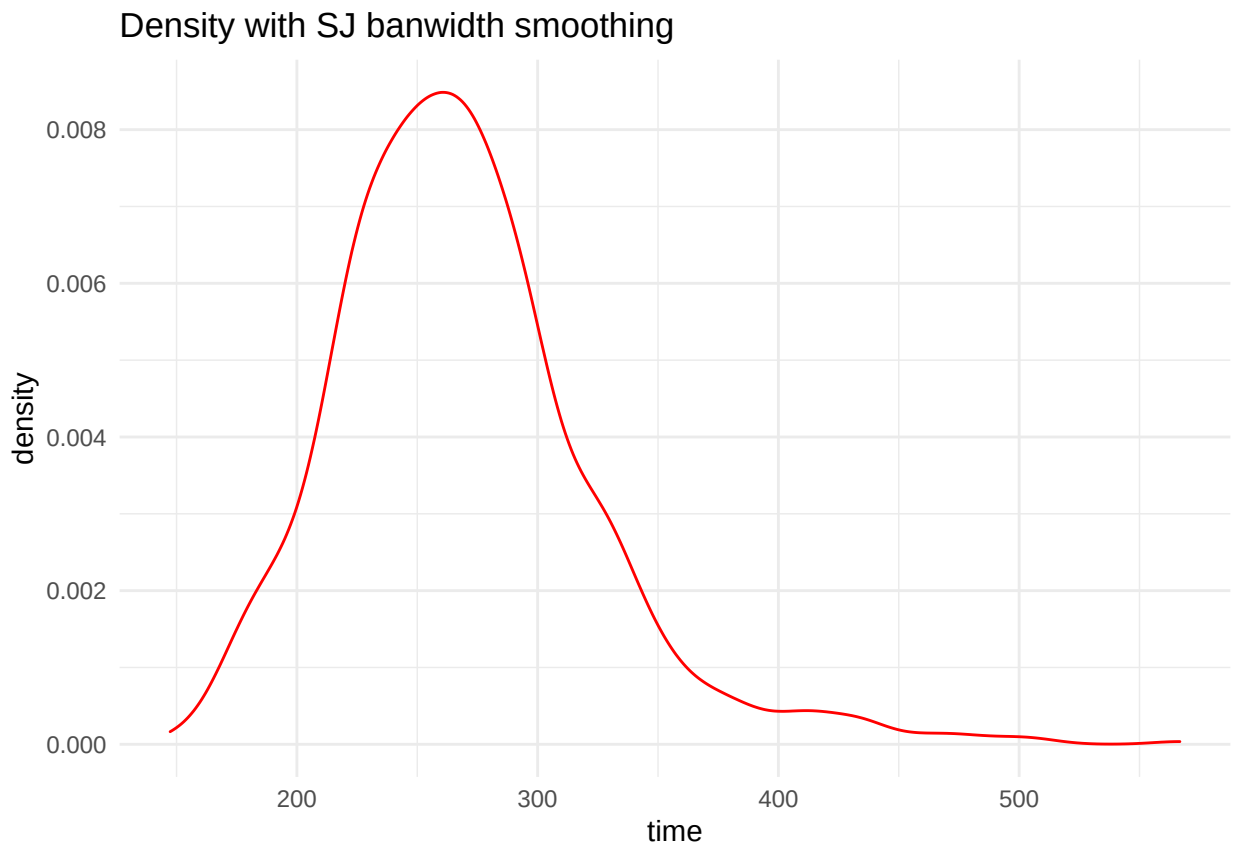
6. New York marathon (R, Home Exercise)

The `nym.2002` dataset (in the UsingR package) contains information about the times taken by participants of the 2002 New York marathon, along with information like age and gender. First of all, bring the dataset into scope by loading the UsingR library: `library(UsingR)`.

a. Plot the kernel density estimate of the `time` column. Given that we have other information available in the dataset, is such a histogram informative? Discuss about this.

Solution:

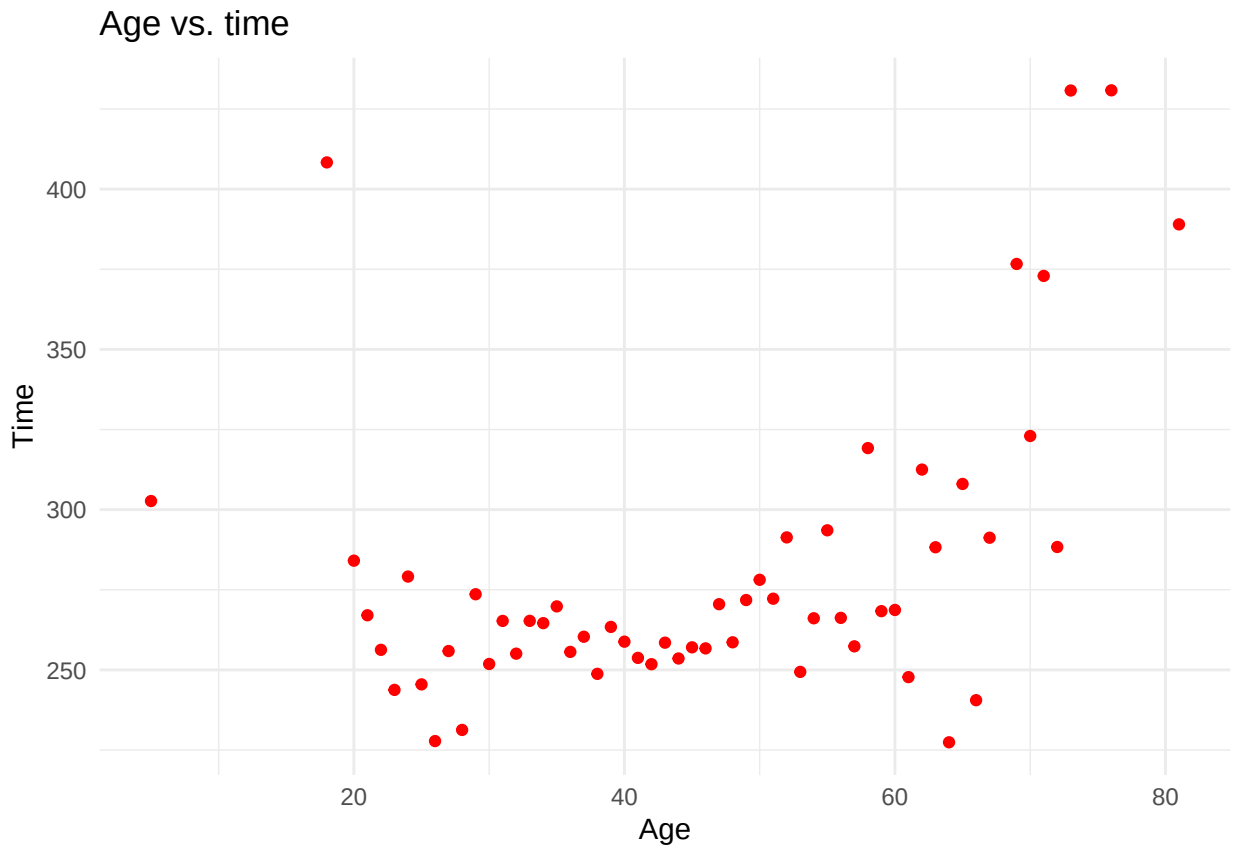
```
ggplot(nym.2002, aes(x = time)) +
  geom_density(bw = 'SJ', color = 'red') + ggtitle("Density with SJ banwidth smoothing") + theme_minimal
```



b. Consider the variable age in combination with `time`. Compute the median time for each age group. To this end, you can use the `aggregate` function, which takes two vectors of the same length (the first one are the values, the other the grouping variables) and a function to aggregate the values belonging to the same group. Build the following plot: on the x axis we have the age, and on the y axis we have the corresponding median time. What do you observe?

Solution:

```
ggplot(nym.2002) +
  stat_summary(aes(x = age, y = time), fun = median, geom = "point", color = "red") +
  labs(title = "Age vs. time", x = "Age", y = "Time") +
  theme_minimal()
```



There is a 5-year-old who finished the marathon in about 300 minutes with no nationality:

```
knitr:: kable(nym.2002 %>% filter(age == min(age)))
```

	place	gender	age	home	time
18260	18589	Male	5		302.6833

c. Plot the kernel density estimate for each age group. To do so, you can use the following tools:

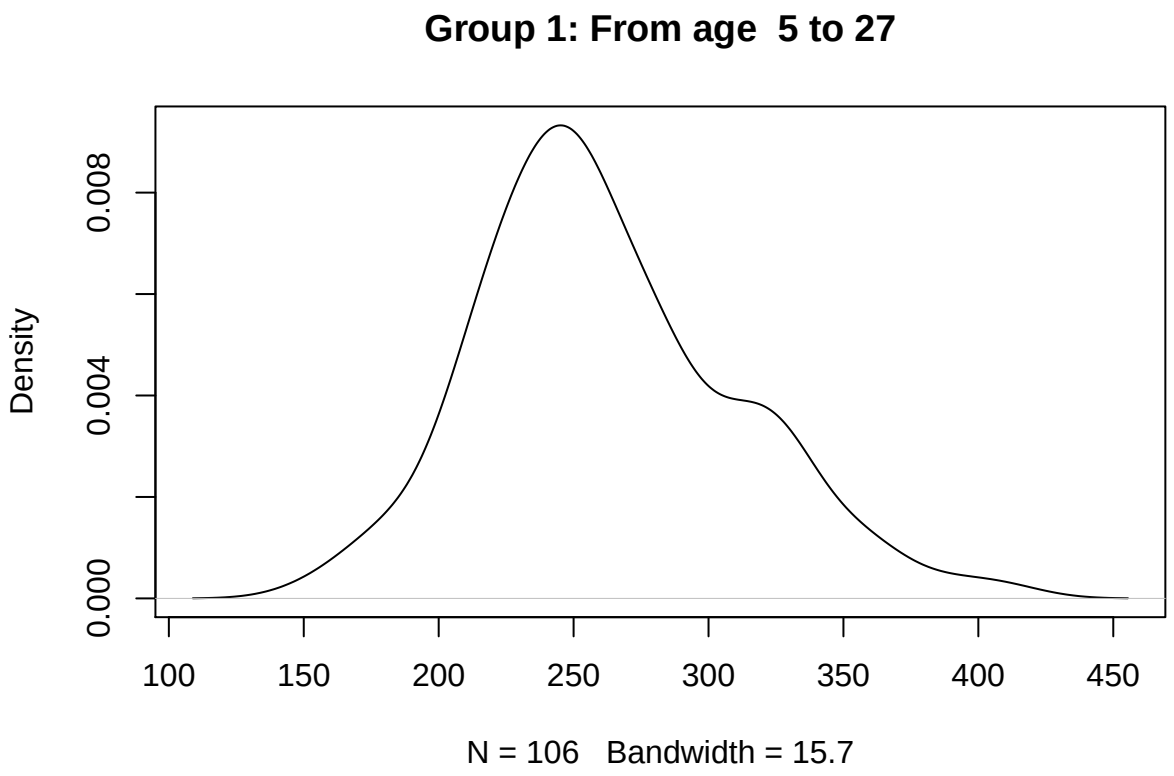
1. Get the set of ages in the dataset using the `unique` function on the `age` column of the dataframe
2. Select the rows corresponding to an age using the `subset` function, which is described on page 159 of Verzani's book. In short, you can use `subset(dataframe, subset = column_name == value)` to select all the rows with the given `value` in the column `column_name`.

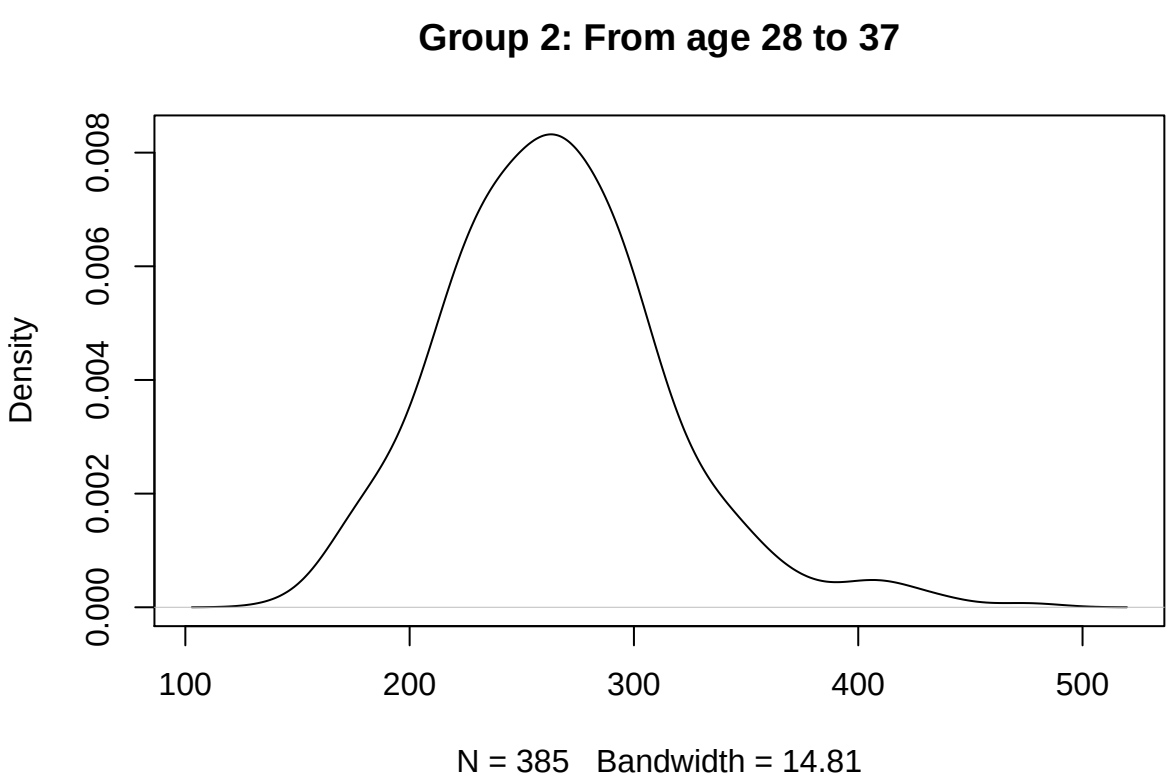
What do you observe in the plots? What might be a possible explanation? Is the median used in the previous point a good summary for each group?

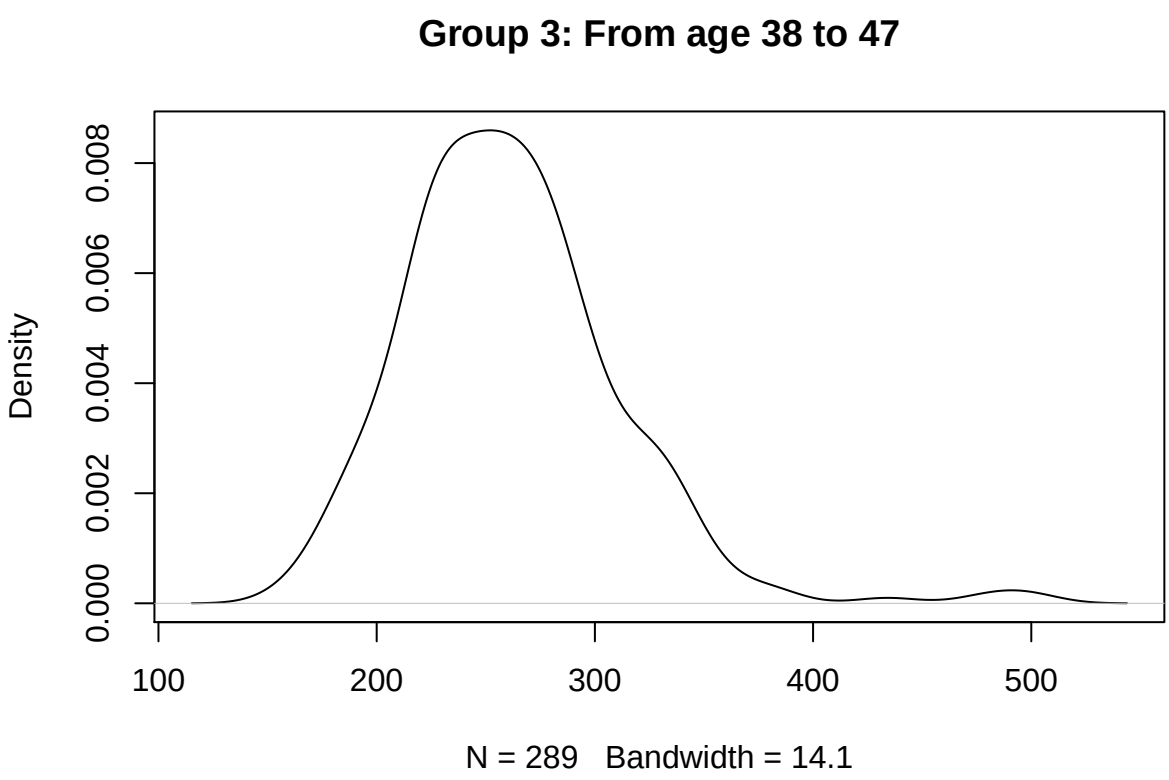
Solution:

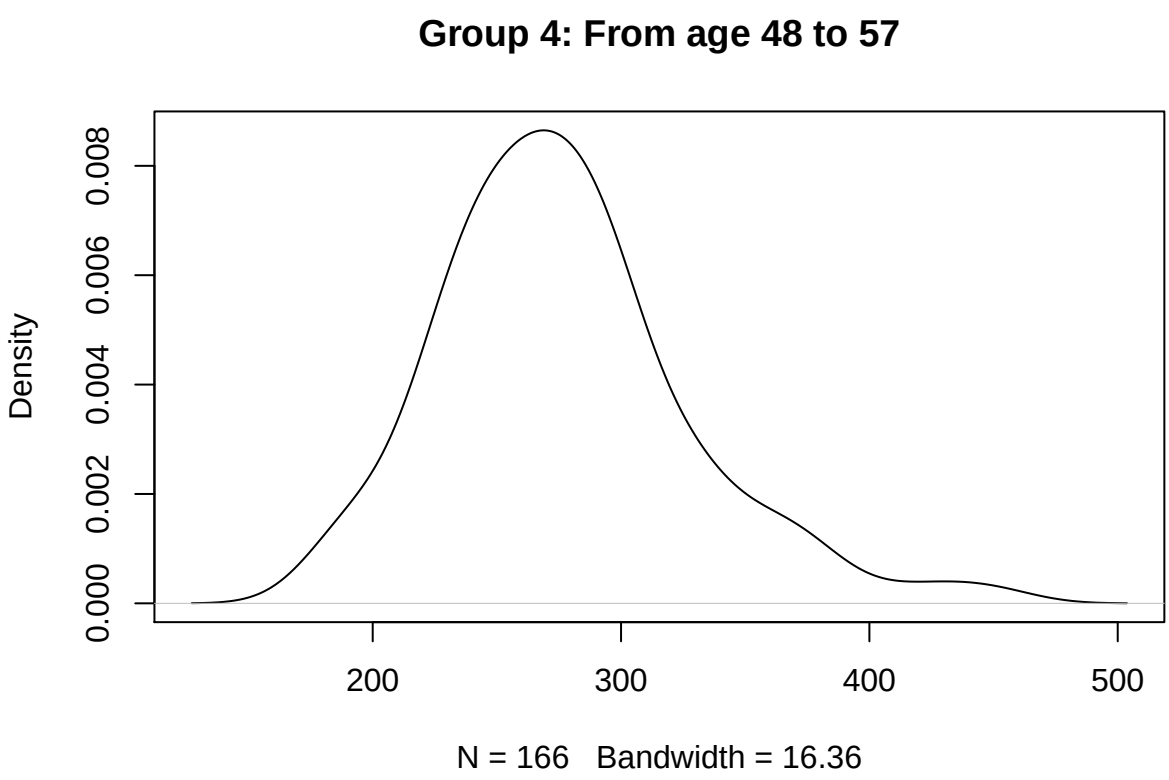
```
ages <- sort(unique(nym.2002$age))
chunk_size <- 10
chunks <- split(ages, ceiling(seq_along(ages)/chunk_size))
plots <- list()

for (i in 1:(length(chunks))) {
  chunk <- unlist(chunks[i])
  s <- subset(nym.2002, age %in% chunk)$time
  title <- sprintf("Group %d: From age %2d to %2d", i, min(chunk), max(chunk))
```

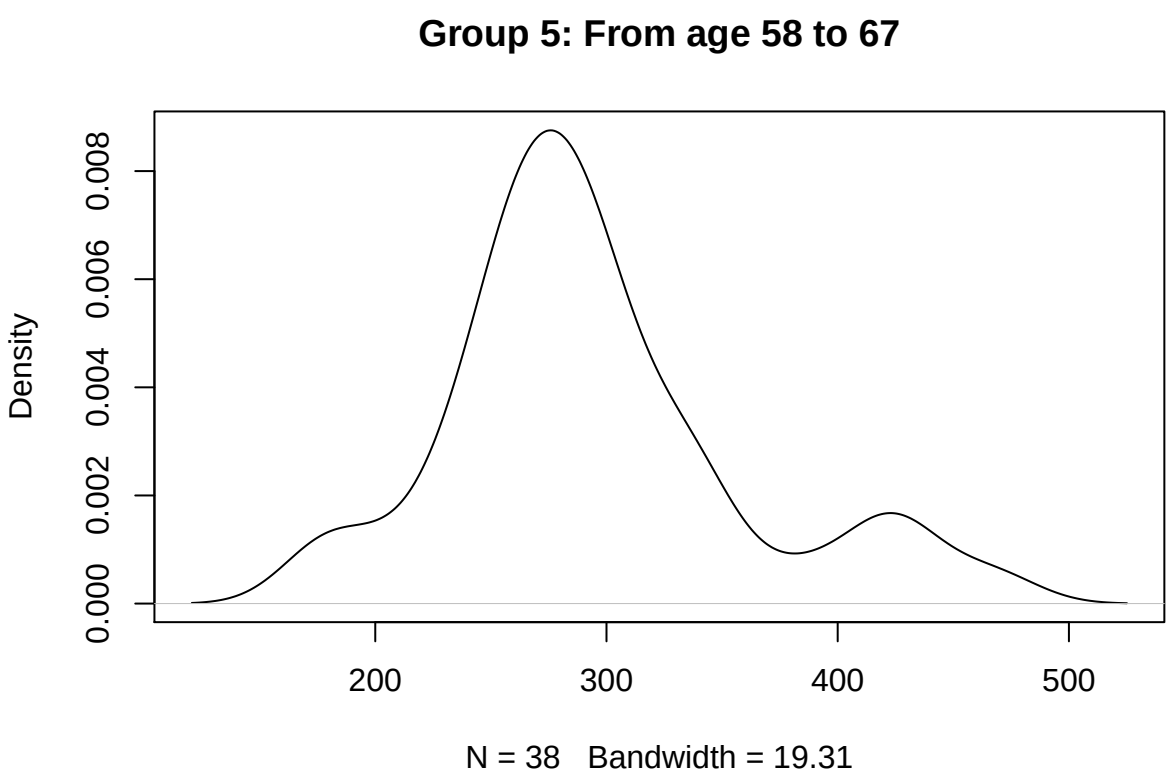
```
  plot(density(s, bw = "SJ"), main = title)
}
```

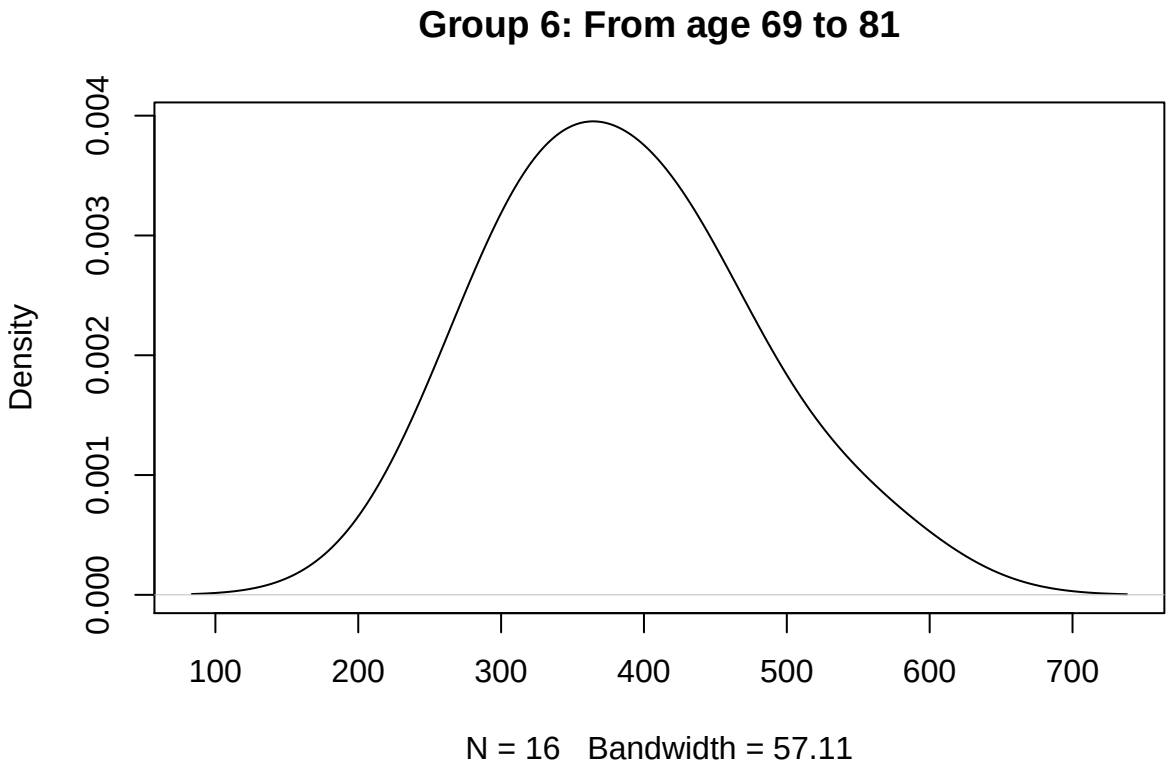



Group 2: From age 28 to 37
Density
0.000
0.002
0.004
0.006
0.008
100
200
300
400
500
N = 385 Bandwidth = 14.81


Group 3: From age 38 to 47
Density
0.000
0.002
0.004
0.006
0.008
100
200
300
400
500
N = 289 Bandwidth = 14.1


Group 4: From age 48 to 57
Density
0.000
0.002
0.004
0.006
0.008
200
300
400
500
N = 166 Bandwidth = 16.36


Group 5: From age 58 to 67
Density
0.000
0.002
0.004
0.006
0.008
200
300
400
500
N = 38 Bandwidth = 19.31


Group 6: From age 69 to 81
Density
0.000
0.001
0.002
0.003
0.004
100
200
300
400
500
600
700
N = 16 Bandwidth = 57.11